Computer-assisted text analysis methodology in the social sciences

Alexa, Melina

Veröffentlichungsversion / Published Version
Arbeitspapier / working paper

ZUMA-Arbeitsbericht 97/07

Computer-assisted text analysis methodology in the social sciences

Melina Alexa

October 1997

ZUMA
Quadrat B2,1
Postfach 12 21 55
D-68072 Mannheim

Telefon: (0621) 1246 - 222
Telefax: (0621) 1246 - 100
E-mail: alexa@zuma-mannheim.de

Computer-assisted text analysis methodology in the social sciences

Melina Alexa

Zentrum für Umfragen, Methoden und Analysen (ZUMA),

B2, 1, PO Box 12 21 55, D-68072 Mannheim, Germany

Email: alexa@zuma-mannheim.de

Abstract

This report presents an account of methods of research in *computer-assisted text analysis* in the social sciences. Rather than to provide a comprehensive enumeration of all computer-assisted text analysis investigations either directly or indirectly related to the social sciences using a quantitative and computer-assisted methodology as their text analytical tool, the aim of this report is to describe the current methodological standpoint of computer-assisted text analysis in the social sciences. This report provides, thus, a description and a discussion of the operations carried out in computer-assisted text analysis investigations.

The report examines both past and well-established as well as some of the current approaches in the field and describes the techniques and the procedures involved. By this means, a first attempt is made toward cataloguing the kinds of supplementary information as well as computational support which are further required to expand the suitability and applicability of the method for the variety of text analysis goals.

1 Introduction

1.1 Text data

Textual data are a rich source of social information. They are an important resource for the social analysts to interpret and describe social behavior, structures, values or norms. Text analysis is significant to social science research and a wide range of techniques has come forth for managing and handling texts, for coding parts of a text, for developing categorization schemes which are used for the coding of texts, for describing and coding the relationships between different text elements and generally for exploring textual data.

What does one mean by textual data? The general and short answer to this question may be in the lines of 'any text which constitutes a relevant and necessary source material for answering the questions one is interested in'. To be more specific the following are all kinds of textual data that can be used for sociological text analysis: open responses to questionnaires, newspaper editorials, commentaries, titles, articles, different kinds of reports, e.g. company annual reports, memos, newspaper reports, etc., journal articles, advertisements, public speeches, conversations, interviews, letters, slogans, keywords, etc.[1]

[1] Of course, text is not the only communication medium which can be subjected to content analysis: images, films, videos, music, dreams are some other media which may be analyzed for content; however, here we restrict ourselves to textual data.

Depending on the aims of text analysis, the textual data may belong to a single semantic domain, e.g. political speeches, or may cut across a number of semantic domains. Furthermore, they may consist of one particular text type, for example only personal letters or articles or speeches, etc., or they may include texts belonging to different text types. A separate factor for compiling a corpus of texts for analysis concerns the speaker/author dimension: the corpus may contain only those texts which have been produced by the same speaker/author - and which may or may not belong to different text types. Alternatively, texts which have been produced by different persons - as in the case of open responses to questions in social survey - may comprise the text corpus. Which texts are relevant for the analysis purposes is determined during the project planning, sampling and data collection phases, which proceed the analysis phase. Sampling is organized around two tasks: define the universe of relevant text communication to be analyzed and decide whether a full or partial coverage should be performed. The universe of relevant communication material relates to the kinds of textual data to be collected and to the kinds of text 'producers': source of texts, text types and speaker or speaker groups (see also section 4.1). Sampling decisions are, then, not only directly related to the purposes of text analysis in order to base text analysis on relevant and adequate material, but also directly affect the analysis results.

What kinds of questions do social scientists ask in relation to the textual material they have collected? The main objective of text analysis within this research field is not different from the main objectives of social sciences research in general, namely to arrive at a specific explanation of social behavior, values, structures or norms. This explanation is directly related - among others - to explaining how this is manifested in spoken and/or written language. Text analysis in social sciences does not aim to provide a description of the linguistic features of texts, which are characteristic of a specific social body, or a complete linguistic text description; such preoccupation in itself is not interesting for the field. Rather by means of text analysis the social scientist has at hand a tool which assists her to either *describe* or *classify* or *interpret* or *make inferences* about social norms or values or behavior or structures based on a corpus of 'real' data, i.e. naturally produced textual data, representative of, as well as relevant to, the particular context(s) of situation to be investigated.

1.2 Text analysis

There exists a large number of approaches to text analysis in the social sciences: for instance, content analysis (for an overview see Weber, 1990 also Krippendorff, 1980, Früh, 1991), concordance analysis (Ellis, 1968 provides a concordance of the works of Shelley), conversational analysis (Sacks, 1972), computational hermeneutics (Mallery, 1985), qualitative text analysis (Kelle, 1995, Weitzman & Miles, 1995), discourse analysis (Polanyi, 1985), linguistic content analysis (Roberts, 1989, 1997a), semantic text grammars (Franzosi, 1987, 1990a, 1990b, 1997), procedural task analysis (van Lehn & Garlick, 1987), map analysis (Carley, 1993), network analysis (Roberts & Popping, 1996), proximity analysis (Danowski, 1988), protocol analysis (Ericsson & Simon, 1984). The advantages of each approach are dependent on the types and amount of texts analyzed and the questions the analyst has set to answer. Even with a single approach variations regarding its application may occur. There is no single technique which is the most appropriate for all kinds of text analysis.

In empirical social sciences a dominant text analysis method has traditionally been content analysis. In fact, content analysis is often used interchangeably with text analysis. Content analysis is an analytical tool, a method, whose usage is not restricted to the social sciences only, but rather its application is broader, for instance, content analysis is used in humanities, psychotherapy, information science, linguistics, etc.

As a social sciences method, content analysis belongs to the general field of empirical research, and empirical text analysis aims at analyzing communication as realized by textual data (as opposed to numerical/statistical data). Content analysis enables quantitative analysis of large numbers of texts in terms of what words or concepts are actually used in the texts.

1.3 Computer-assisted text analysis

Computer-assisted content analysis is a special instance of text, and in particular of content, analysis methodology in the social sciences. According to Zuell *et al.* (1991) *"Die cui (computerunterstützte Inhaltsanalyse) umfaßt dann innerhalb der Inhaltsanalyse alle Verfahren, bei denen die Zuordnung von Textmerkmalen zu Variablen mittels Algorithmen, d.h. eindeutig festgelegten logischen oder statistischen Operationen, geschieht."* (p. 17) [2]. Zuell *et al.* (1991) thus see content analysis as a method which allocates scheme categories to text, based on defined variables. Variables are indicators of both manifest, e.g. author, social context, etc., and latent constructs. The latter are constructs which are not already given or cannot be directly extracted, for example, positive, negative or authoritarian statements. Such variables can be combined with variables of other methods in order to build up complex models.

At present, and in contrast to the sixties and seventies when availability of electronic text was minimal and the conversion of existing text in machine-readable form a complex and time-intensive process, access to machine-readable text is fairly straightforward, and considerably less complex and laborious a process. In fact, we are experiencing the benefits (and to a certain extent the frustrations) of available machine-readable text in abundance. Nowadays, there exists a large number of electronic text archives containing text data from a large variety of sources and for various purposes. Moreover, there is a proliferation of on-line text databases with a wide variety of text material which the analyst can directly access and download for her own research purposes. In addition to these, there are available text corpora in various languages, full texts of a variety of publications are made available on-line, there exist numerous archives of the electronic communication of a variety of discussion groups all of which improve the researchers' access to text material they are already in machine-readable form and can in a relatively straightforward manner be used for computer-assisted text analysis purposes. The possibility to convert printed material in electronic form fairly fast by using optical scanner readers, although not only as reliably as one would have hoped for, has increased also the degree of easiness with which the analyst can come into electronic text material. The suggestions and methods proposed by DeWeese (1976, 1977) for obtaining machine-readable text a couple of decades ago may nowadays be considered as belonging to the 'history' of the first steps of computer-assisted content analysis.

With both the increased availability and ease of accessing electronic text computer-assisted, computer-assisted text analysis becomes steadily important for discovering meaning. This is coupled with the availability and, to a some extent, the increase of user-friendliness of a variety of software programs which support text analysis. Furthermore, the number of application contexts of computer-assisted text and content analysis has increased, ranging from educational purposes (Eltinge & Roberts, 1993, Eltinge,1997) to multimedia applications (see for instance the work of the Movie Content Analysis (MoCA) group, Lienhart *et al.* 1996, Fischer *et al.* 1995) and information retrieval tasks (Siemens, 1994). These are factors which have significant consequences for text analysis in general: different techniques are emerging or old ones are put into practice, tested and their weaknesses are discovered. Moreover, ways to improve or refine established techniques or,

[2] English translation: Computer-assisted content analysis covers all content analysis steps for allocating text features (characteristics) to variables by means of algorithms, that is, explicit, unambiguous logical or statistical operations.

generally, bring these in tune with the current technological possibilities are proposed, implemented and tested.

Recent methods of approach in computer-assisted text and content analysis (see sections 5.4 to 5.8) attempt to improve efficiency as well as answer complex questions by incorporating layers of different kinds of meta-information or different kinds of linguistic description. The existing model of assigning properties (categories), which are heuristic rather than conceptually based, to word forms and counting frequencies of occurrence of these properties is alone not sufficient for the variety of research questions and application contexts of text analysis. Furthermore, it has been often demonstrated that content cannot be analyzed without taking into account the general context of situation a text belongs to. Encompassing, thus, insights from linguistics with regards to contextual and discourse analysis should benefit the traditional model of computer-assisted content analysis.

Attempts to enhance computer-assisted content analysis by incorporating linguistic information and applying natural language processing techniques are reported in Roberts & Popping (1996), McTavish *et al.* (1997), Nazarenko *et al.* (1995), Wilson & Rayson (1993), Franzosi (1987, 1990a, 1990b, 1997) and Roberts (1989, 1997a). They all involve different aspects of either using or stressing the advantages of using linguistic knowledge for gaining more general and generally *more* information for analysis purposes.

1.4 Aims of the report

This report has two complementary objectives: First, it provides a critical account of the approaches to computer-assisted text analysis within the social sciences context. Since computer-assisted content analysis is a particular instance of computer-assisted text analysis which has dominated methodologically the social sciences context, a short background to the method is provided and some of its key concepts are explained with the purpose to elucidate both concepts and methodological decisions discussed in the recent approaches.

The second objective of this report is to attempt a first assessment of the current methodological standpoint of computer-assisted text analysis in the social sciences. This assessment should serve as the basis for determining the kinds of computational requirements for computer-assisted text analysis methodology in order to support a wide spectrum of investigations.

1.5 Overview

A brief background to computer-assisted content analysis is provided in the next section together with some definition of its key issues and aims. Section 3 discusses the relation of content analysis to general text analysis and elaborates on the possible contributions linguistics can make to computer -assisted content analysis for the social sciences. Section 4 highlights the basic phases of computer-assisted analysis methodology. A review of methods of approach is provided in section 5, which discusses critically well-established as well as new methods. The final section of the report summarizes the main issues for the future development of computer-assisted text analysis methodology and the required software to support it which are raised by the variety of the reported approaches.

2 Computer-assisted content analysis - some background

Historically, ***content analysis*** has been mainly associated with research in journalism. This research influenced early empirical political science, which strengthened content analysis methodologically. As reported in Stone *et al.* (1966, pp. 24-25) *"The work of Lasswell, Leites, and associates in the 1940's put both content analysis theory and method in a much more refreshing and meaningful perspective. Based on an extensive background in the study of propaganda (Lasswell, 1927), their work at the University of Chicago and at the Experimental Division for the Study of Wartime Communications at the Library of Congress offered a major opportunity to make advances both in conceptualization and technique, summarized in their Language of Politics (1949)."*[3] Although content analysis practice can be traced back to the previous century with the analysis of Zion hymns (see Dovring 1954), published groundwork on its theoretical basis can be placed in the time after the second world war with the work of Lasswell *et al.* (1952), Lazarsfeld & Barton (1951), Berelson 1952, or the published results of the Allerton House Conference in 1955, edited by Ithiel de Sola Pool (Pool, 1959), which attracted the most important content analysis researchers of that time.

Work on the theory operationalization or building strategies by Osgood, Lasswell, and others took place in the sixties, which is also the time of the first significant steps towards the development, usage and exploitation of computers for the purposes of assisting the content analysis of texts, in other words towards ***computer-assisted content analysis***. Programs such as the General Inquirer (Stone *et al*. 1966, Zuell *et al.* 1989) or WORDS (Iker & Harway 1969) are products of that time. The approaches to computer-assisted content and text analysis and, the methodological considerations which have been developed during the sixties and the key concepts, work and considerations were presented in the book edited by Gerbner *et al.* (1969), which was based on the Annenberg School content analysis conference, which took place at Philadelphia in 1967.

A perusal of the approaches reported in the Annenberg School conference book (Gerbner *et el.*, 1969) evidences a contrast between those approaches emphasizing theory operationalization and whose researchers are interested in using computers to categorize texts by means of coding, according to a predefined and pre-constructed categorization scheme, those approaches whose proponents are interested in statistical techniques, for instance, in factor analyzing word counts to empirically extract the recurrent themes of their text material, and thus are not based on a specific theory, and those researchers who saw in computerized analysis a potential for a more effective management and organization of their textual information.

During the seventies progress on the computer-assisted analysis front is slow. Computer technology is still not broadly available, not user-friendly, neither affordable nor widely accessible for the majority of investigators in the social sciences. For the most of the seventies and during the largest part of the eighties a shift of importance appears to have taken place in content analysis. Application-oriented work was brought in the foreground, in other words practical research as opposed to 'grand theories'. The recent bibliography on text analysis in the social sciences text analysis shows the focus of attention being on developing and applying a text analysis methodology which is practical and comprehensive enough for the content analyst to make valid inferences or comparisons or produce valid descriptions (see for example Früh 1991, Roberts, 1997b, Kabanoff, 1996, McTavish & Pirro, 1990, Geis & Zuell, 1996 among others.

---

[3] Lasswell, H. D. (1927): *Propaganda Technique in the World War*. Knopf, New York.

Language of Politics was reprinted in 1965 by MIT Press: Lasswell, H. D., N. Leites and associates (1965): *Language of Politics: Studies in Quantitative Semantics*. Revised Edition. MIT Press, Cambridge, Mass.

As mentioned in the introduction, recent methods of approach attempt to incorporate linguistic information and account for the relational properties of analyzed concepts. There is an interest in using linguistic resources and description and integrating artificial intelligence and natural language processing techniques with computer-assisted content analysis. Some of this work will be reviewed later on in this report. It is necessary to note that this interest comes as a consequence of a more intensified and wider application of computer-assisted content analysis, given on the one hand the technology 'maturity' and on the other hand the availability of electronic text and means for managing it.

To summarize: although during the sixties effort and optimism was invested in computer-assisted content analysis as a text analysis instrument, this was not followed on in the seventies, mainly due to the fact that computer technology was still in main development as well as that computer access was not broadly possible. Added to that was the fact that machine-readable texts were scarce and the conversion of written or spoken text into electronic form required great effort.

In the mid-eighties and early nineties progress has been made towards the development of computer-assisted text analysis software to support a variety of analysis tasks as well as different languages, enabling the application of content analysis as a method of text analysis. A variety of software for both qualitative and quantitative computer-assisted analysis has appeared during this time: qualitative analysts have come to recognize the advantages of computer technology to support theory development as well as efficient organizing of their interpretation and qualitative software (see Weitzman & Miles, 1995, Tesch, 1990) assist in theory building and hypothesis testing, enabling researchers to mark up (or code) texts, but they are not intended to be used for automatic coding. Different to qualitative, quantitative-oriented software has a longer tradition, but during that time it has mainly invested on technology and improving user-friendliness with methodological issues becoming secondary.

2.1 Modes of content analysis: automatic, manual, integration of computerized and manual coding

Traditionally, content analysis has been distinguished between computer-assisted and manual content analysis. The main distinction here lies in how the coding is performed: in computer-assisted analysis the goal is to *automatically* (without human intervention) code specific parts of text - mainly words - according to a particular categorization scheme. In manual content analysis (Früh, 1991) coding is not performed automatically; instead a group of human coders performs *manually* the required coding.

The general research scope and context of content analysis as well as practical aspects, such as time, costs and experience determine which mode is more appropriate for a given project. A comprehensive and well worked out check-list to assist the analyst to make the right methodological decision is provided in Geis (1992). To mention some of the advantages and disadvantages of computer-assisted analysis in the sense of automatic coding, one disadvantage, given that coding of words is performed without taking into consideration the context in which they occur, is that automatic coding can go 'terribly' wrong. For that reason word-sense disambiguation procedures are necessary and work on defining rules to disambiguate homographs, i.e. words like *bit*, *like*, *kind*, etc. having more than one meaning, is reported as early as in the sixties (see for instance Stone, 1969).

Additionally, the kind of texts which are best analyzed/coded automatically are those which are relatively restricted semantically. A principal advantage of automatic coding relates to the quantity of the text to be analyzed; for very large text corpora one can hardly expect that these can

be coded within a realistic time schedule and budget without automated coding procedures. Furthermore, the coding is replicable and explicit, there is no inter-coder variation, plus a number of pre-tests as well as categorization scheme refinement and modification are possible.

Nowadays, the traditional distinction between automatic and manual or computer-assisted and coder-based is steadily becoming blurry. The existence or non-existence of supporting user-friendly and accessible technology is no longer a constraint for the researchers to use a single mode of analysis.

With the development and application of software programs and general computing environments for content analysis another mode has emerged, namely that of integrating human and computerized coding. Programs such as the PLCA (Program for Linguistic Content Analysis) which supports Roberts' method of linguistic content analysis (Roberts, 1989, 1997a) or the MECA (Map Extraction, Comparison, and Analysis) suite of programs which support the map analysis approach (Carley, 1988, 1993, Carley & Palmquist, 1992) are examples of computer-assisted content analysis, where coding integrates automatic and on-line coding by coders.

We distinguish here between the above computer-assisted approaches and those ones, which, although termed computer-assisted (for instance, the computer-assisted analysis work reported in Franzosi, 1987, 1990a), the coding is performed by human coders in an interactive, on-line mode with the computer recording the coding. We see these approaches as categorized under the manual mode of content analysis and not under automatic or integrative modes.

A final point to make, concerns the research reporting on either the assessment of the effectiveness of this coding mode, e.g. Franzosi (1990b, 1995), or on increased reliability and replicapility, e.g. Carley (1988) , or on comparing the manual and computer-assisted coding, e.g. Morris (1994), Hohnloser *et al.* (1996). Such work provides valuable information to the content analysis researcher for choosing the appropriate analysis mode.

2.2 Frequency of occurrence: word, lemma, scheme category

In computer-assisted content analysis, the frequency of a word form as it appears in a text corpus is a predominant source of information. The frequency of occurrence of a word is used to indicate the important or recurrent themes in a text or a corpus of texts and it may form the basis for empirically determining scheme categories and dictionary entries (see section 4.3 below). Lemmatization, whereby different word forms are grouped under one lemma, may be applied and the frequencies of lemmas may then be taken into account. Often, however, it is the overall frequency of occurrence of the word form and not of the lemma which is used for analysis.

Furthermore, the single word form constitutes typically the coded instance, the latter being the basis of interpretation. Multi-word phrases or idioms are considered as one unit only if they have been entered as multi-word forms in the dictionary used for coding. Consequently, the frequency of occurrence of scheme categories is based on single words which have been coded according to the scheme categories.

The dependence on or importance of the frequency of occurrence of words has been criticized as a deficiency by a number of researchers since the early days of computer-assisted content analysis: Goldhamer (1969), for instance, argued that *"In the past, content analysis has been restricted to word counting and closely related operations, although it has been clear that treatment of the overall, broad meaning of a text is a prerequisite for its satisfactory analysis. This conclusion has arisen from the failure of computerized content analysis to deal with aspects of natural language such as ambiguity, metaphor and humor. We ascribe this failure to the fact that computerized content analysis has attempted to*

go directly from the text to the investigator's set of categories without taking into account the mediation of language--of broad meanings embedded in the language and culture as a whole." (p. 347).

Hays (1969) has criticized the theoretical preoccupation for accounting only for frequencies: *"Theories that account only for frequencies of certain words or classes of words or for the association of certain combinations of words are weak (that is, poor) theories."* (p. 65). He dismisses the model of assigning properties (categories) to words and counting frequencies of occurrence of these properties as insufficient: *"The content of a message should be analyzed in terms of what has gone before, in terms of the relationship between the message and what the sender and receiver already knows."* (Hays, 1969, p. 66).

One should consider, however, the general research scope and aims of analysis: one should not be too hasty to dismiss a method that may provide valid information about recurring themes in a large body of electronic textual material if that is in fact the purpose of the analysis project. For instance, the approach suggested by Iker & Harway (1969) for empirically recognizing and analyzing major content themes of a large amount of electronic text (see section 5.2), selects for analysis a subset of the total of the corpus types based on their high frequency of occurrence. Iker (1974) acknowledges that this is not necessarily to be interpreted as the sole criterion for selecting such subset: instead he states that *"A frequency criterion, then, is needed but only to set a lower bound below which words are neither evaluated nor selected.."* (p. 314). Although a frequency criterion ensures the inclusion of high-frequency words as well as the exclusion of low-frequency words, if one is interested in identifying how words are associated then high frequency words have no information about how they associate with others: *"Indeed, words with **very** high frequency tend to be associationally impoverished; if a words always occurs, i.e. appears in all segments, and does so with a relatively flat frequency, it tends to relate poorly to most other words since there is no way to contrast its "ups and downs" with those of other words. On the other hand, words with very low frequencies need to be avoided since they involve potentially unstable relationships."* (Iker, 1974, p. 314).

Whereas Hays (1969) is interested in incorporating semantic and contextual knowledge for analysis, dismissing as weak analysis based on frequency, Iker (1974) demonstrates how frequency information can be utilized for particular analysis aims. Clearly, different research questions and application contexts require or motivate different analysis techniques which may or may not profit from frequency information. Whereas for a specific analysis within a particular research context frequency information might be of primary importance, it might be a poor indicator for another.

2.3 Qualitative and quantitative analysis

During the earlier days of content analysis Lasswell *et al.* (1952) declared that *"There is clearly no reason for content analysis unless the question one wants answered is quantitative"* (p. 45). Berelson's (probably most) quoted definition of content analysis states that *"Content analysis is a research technique for the objective, systematic, and quantitative description of the manifest content of communication"*(Berelson 1952, p. 18). Also Cartwright (1953) defining content analysis states that *"We propose to use the terms "content analysis" and "coding" interchangeably to refer to the objective, systematic, and quantitative description of any symbolic behavior"* (p. 424).

The quantitative requirement for content analysis has stressed the importance and relevance to analysis of the frequency with which words or themes appear in a body of texts. It has been clearly demonstrated by Holsti (1969, pp. 7-9) that restricting content analysis to frequency counts presents theoretical and practical problems and that measures other than frequency may be useful. The term quantitative may take on many meanings, no one of which will be most suitable for every type of research.

As a counter requirement to quantitative, qualitative analysis is an act of interpretation which is not based on measurement of frequencies of occurrences or statistical significance tests. Although the quantitative-qualitative distinction is not a trivial one, in computer-assisted analysis its rigidity may appear rather artificial in nature. It is well accepted that rigorous quantitative studies may use non-numerical procedures at various stages in the research such as the initial selection of categories, or check face validity of the quantitative results after coding has been performed. Besides quantitative results may highlight qualitative aspects of the text which might have escaped otherwise the researcher's scrutiny. In that sense the requirement for a qualitative as opposed to qualitative analysis appears to be a 'shadow' dichotomy. In fact, rather than dichotomous attributes, quantitative and qualitative fall along a continuum (Lazarsfeld & Barton, 1951) instead. It was noted by Holsti in the end of sixties that *"the quantitative requirement has often been cited as essential to content analysis, both by those who praise the technique as more scientific than other methods of documentary analysis and by those who are most critical of content analysis."* (Holsti 1969, p. 5). He suggested, however, and this appeared to be accepted until the time of his writing by the research community that *"the content analysts should use qualitative and quantitative methods to supplement each other: It is by moving back and forth between these approaches that the investigator is most likely to gain insight into the meaning of his data"* (Holsti, 1969, p. 11).

This complementarity between quantitative and qualitative analysis is by now accepted, for instance see Früh (1991), McTavish *et al.* (1995), Kabanoff (1996), among others, or encouraged as in Evans (1996). To give an example, Kabanoff (1996) talking about the advantages of text analysis (in his article text analysis is shortened to TA) for organizational behavior research states that: *"One of the benefits of TA is that it combines desirable characteristics from what are generally considered two separate, even inimical research traditions - qualitative and quantitative research. By allowing us to deal systematically with, and to quantify what are normally considered qualitative data such as documents and interviews, TA helps address the criticisms of sterility and lack of relevance that are sometimes directed at traditional, quantitative forms of research. At the same time, TA permits those of us who believe that, in the end being able to quantify a phenomenon is desirable, to convert qualitative data to a quantitative form."* (p. 5). It is important to stress here that computer-assisted analysis may support the combination of the two approaches.

As a final point of clarification, it may be suggested that the emphasis on the quantitative aspect of analysis relates more to the fact that content analysis belongs to the empirical research tradition and by this one refers to the provable, systematic, objective, data-based approach, rather to the pure counting of words that it has often been accused for. If quantitative, as a concept or feature, is at all to be used for the definition of content analysis, then it should not mean the mere counting of words or relevant scheme categories, but rather as put by Kriz (1978) *"(quantitativ) bedeutet gerade die Möglichkeit, komplexe Strukturen hinreichend einfach intersubjektiv zu beschreiben - vergleichbar der Abbildung orchestraler akustischer Klangkomplexe mittels der Noten in einer Partitur."* (p. 49).[4]

3 The relation of content analysis to text analysis

Text analysis is a general field of practice defined within and employed by a large variety of disciplinary fields. With regards to content analysis, text analysis is the general or overarching field of activity. The purposes of text analysis differ depending on discipline. The difference in purposes influences the methodology used. Content analysis is viewed either as a method or a theory or a set of techniques for providing interpretations of texts (for the latter see Deese, 1969). This may explain

[4] English translation: quantitative means exactly the possibility, to adequately describe complex structures in an objective manner - comparable to the illustration of orchestral acoustic music complexes in the middle of the notes of a music score.

the non-agreement about the analysis output as well the misunderstanding one can observe between different researchers within one discipline or across disciplinary fields. Mayntz *et al.* (1978) argue that different from the linguistic text analysis, in content analysis one's aim is to discover and categorize the content or *meaning* of certain linguistic configurations - words, word combinations, sentences, etc. Of course, other disciplinary fields (linguistics, psychology, literary criticism) also define meaning extraction and categorization as their aim. The difference lies in the interpretation framework for analysis. For example, in functional linguistics a specific aim of text analysis is to provide a categorization of noun phrases according to a number of functional categories, such as 'agent', 'medium', etc. For political scientists engaged in text analysis noun phrases - if at all categorized or analyzed as noun phrases - become interesting after they have been categorized as prosecutors, accused, victims, etc.

3.1 The 'meaning' of text in content analysis

A great amount of effort and time has been invested in analyzing texts (of a variety of text types) in the social sciences. Nevertheless very little has been written on how text is to be defined in the field. One of the exceptions is Goldhamer (1969), who provides a definition for *text*, *document* and *word* within the General Inquirer approach.

Kripppendorff (1969) explaining the analytical problem of content observes that "*the analytical problem is to* ***inferentially link available observations****, the* ***raw data, or, in short, text,*** *to specific events, behavior, or phenomena associated with the source*" (my emphasis). In his view, 'available observations', 'raw data' and 'text' denote the same. He explains in a footnote that text is not meant in the restricted sense of it referring to "*linguistic expressions in written form*", but rather he proposes a more general notion of text: "*the term may refer to a variety of data the "texture" of which becomes apparent through its identifiable and distinguishable constituents.*" (p. 8). Although some consideration has already been given into the nature of the data the notion of texture remains nevertheless vague.

One uses content analysis to analyze what is referred to as "communications messages"(Barcus 1959), "messages" (Holsti 1969), "texts", "records", "data" or "content data", to give some examples. All these may involve such texts as interviews, editorials, advertisements, etc. No systematic attention is paid into what kind of consequences these particular 'texts' have for the analysis with regards to the linguistic features which characterize them as particular instantiations of texts. In fact, it is unclear whether the consideration of whether there are particular consequences due to the kinds of texts selected for analysis is made at all. Some analysts appear to be conscious of this failing: Goldhamer (1969) observed that an issue which "*relates to the availability and use of adequate information about the text and its context"* and points that "*decoding text is so subtle a task that information about language and information about the contemporary world of the source are required*"for the performance of computerized content analysis programs. Linguistic information as information about the particular context of situation analyzed and the domain the text belongs to are meant to be important; however, in computer-assisted content analysis in social sciences no information is provided about whether these types of information have been considered during analysis or in what concrete way(s) they have been exploited, for instance what their role has been in the definition of the categorization scheme or the categorization process itself.

Admittedly, considerations about what kinds of texts are to constitute the basis of text analysis are related to the sampling phase; of course, analysis of texts follows the sampling phase. Although the reporting of text analysis results always includes information about the specific kinds of texts analysis has been based on, for example as editorials, comics, interviews, etc., in general, the concept 'text' is used to cover all these, that is source textual data, different text types and each single text. Moreover, although the information about the broad text types (if more than one text

type are analyzed) is recorded and provided, differences between the distinct kinds of language material collected which are due to the texture or text characteristics of this material are not taken into consideration, or if indeed they are, both selection criteria and process are not transparent.

In general, the notion of text as a particular unit of analysis with its own specific features which depend on semantic domain, speaker/writer, lexical choice, text structure devices - all aspects of investigation in linguistics - is missing. Often the concepts 'data' and 'text' are used as synonyms; the fact that the 'text data' may consist of a single text or a number of texts, which perhaps belong to different text types or semantic domains, that the structure of the texts may vary, all factors which influence the language used does not appear to play a significant role for analysis or at least it is not reported as significant. It is nevertheless certain that different texts, depending on the domain and text type they belong to have different characteristics which are represented in the choice of the vocabulary, intonation, grammatical structures, cohesion devices, etc. And these distinctions contribute to the value of text interpretation.

Furthermore, if one is to construct one's own categorization scheme for analysis (as opposed to using an already constructed scheme or dictionary) it is acknowledged that the types of texts which are collected for analysis play a crucial role for the development of the categorization scheme. However, both decisions related to specific text features of the selected data which play a crucial role for the construction of a categorization scheme and those which concern the testing and modification or refinement of a scheme, are rarely described in the literature and, therefore, do not appear significant for the presentation of one's analysis methodology and results. With regards to the analysis or coding per se, one cannot determine whether specific text features have been taken into account and, if so, what kind of consequences they have for the analysis results (for an exception see Franzosi, 1995, and section 5.8). In other words, one can observe a lack of transparency with regards to text specific information in the presentation of specific analyses or methodologies of performed analyses.

3.2 Content analysis and linguistic theories

Broadly - albeit informally - stated the purpose of content analysis is to "dig out meaning". The meaning that one attributes to a text depends on one's theory. To try to answer in a generally acceptable - in the sense of cross-disciplinary - manner the question of what a text or part of a text really means is impossible not to mention uninteresting for research purposes. How meaning is defined is dependent on one's theoretical frame of reference. This, in turn, is what determines the categorization scheme to be used for analysis. To use again linguistics as our example, analysis of a sentence involves the determination of the kind of process the verb can be categorized to, e.g. as 'mental' (e.g. to think) or 'material' process (e.g. to build), whether the noun phrase being the subject of the sentence is an 'agent', an 'actor' or 'medium'. The social scientist using content analysis, on the other hand, makes the kinds of classifications she believes are most useful to her, i.e. those which are relevant to her theory and the hypotheses she intends to test and those of practical use in analyzing the kinds of texts in which she is interested. The linguistic analysis is obviously different from the categorization involved in the social sciences analysis, where such categories may be involved as, for example 'aggression', 'positive', 'success', 'time', where part of speech categorization may be insignificant and where nouns, verbs and adjectives may be lumped together in a single category denoting, for example, success.

In the early fifties, Lasswell *et al.* argued that "*There is **yet** no good theory of symbolic communication by which to predict how given values, attitudes or ideologies will be expressed to manifest symbols. The extant theories tend to deal with values, attitudes, and ideologies as the ultimate units, not with the symbolic atoms of which they are composed. **There is almost no theory of language which predicts the specific words one will emit in the course of expressing the contents of his thoughts**."* (Lasswell *et*

al., 1952, p. 49, emphasis in bold my own). A bit less than 15 years later Philip Stone and his colleagues commented that "*Unfortunately, although our scientific theories may lead us to the concepts we wish to study, we lack an adequate theory of language to direct us in finding the alternative signs that express a particular concept. In a situation where something is to be said**, there is no theory to tell us what words will be used to say it. (...)** To be useful such a theory will have to combine specific knowledge of the individual speaker and the perceived social situation, together with a general knowledge of language.*"(Stone *et al.*, 1966, p. 10, emphasis my own).

It is doubtful - if at all desirable - that a theory of language will get to the point to provide falsesafe rules about all the exact words which may be used to express a certain thing. Moreover, if we accept that language is not static but instead it develops with time, and that words which are used to express particular concepts at a certain time point cease to mean the same or are not used at all or are attached with different connotations, that new words are constantly coined or borrowed and added in the language vocabulary, then this expectation appears unrealistic. A realistic expectation instead, is to have a language theory which provides us with the parameters or the 'tools' to discover and describe the different ways of expressing the same thing according to such factors as (among others) mode of communication, interpersonal aspects and choice of text type. That is, a theory which can provide an account of how (a) language functions and describes its potential for realizing meaning.

In what concrete ways can then theoretical linguistics and natural language processing techniques contribute to computer-assisted the text analysis in social sciences? On the one hand, they can offer a general and 'neutral' meta-language as a basis for further analysis and coding. On the other hand, they can provide surface analysis - part of speech tagging, noun phrase categorization or extraction, event extraction and categorization, complement argument structures, all of which may be used for further processing to support the social scientists' interpretation and description purposes; natural language processing techniques can also offer guidance and experience for dictionary construction (see relevant work on indexing, thesaurus development, machine readable dictionaries, semantic nets) as well as a framework for discovering and describing functional (semantic) information. They can offer already constructed general-language lexical resources to complement such resources as categorization schemes and dictionaries which have been constructed specifically for the analysis purposes of a sociological investigation. All in all, computational linguistics work as for example in computational lexicography, information extraction or parsing can be advantageous for social scientists performing content analysis. It must be stressed, however, that language theories and natural language processing methods and programs are not intended to do 'meaning analysis' as defined in content analysis.

4 Computer-assisted analysis methodology

Two ideal types may be distinguished in computer-assisted content analysis: one is the general, 'a-priori' type (Zuell *et al.*, 1991, p. 15), which relates to the automatic coding based on a categorization scheme (dictionary) constructed by the analyst according to the theoretical framework she is working in. The second, the empirical attempt (Zuell *et al.*, 1991, p. 15), concerns the automatic (inductive) extraction of 'categories' which are indicative of particular topics or themes dealt within a specific body of text data, and which may then comprise the scheme categories to be used for coding. Such automatic extraction is based on the frequency of occurrence of 'content' words and their statistical analysis to determine which words can indicate the categories which should be used for analysis. Both approaches, however, embrace the phases of data collection, transcribing, sampling, preparation for processing, the coding itself, validity and reliability tests for the analysis results and various statistics procedures for the interpretation of the coded data. The following sections discuss some of these phases.

4.1 Sampling: text data selection

Given a well-defined research project, the hypothesis(es) it poses and its objective(s), the first methodological consideration concerns the text material to be analyzed; this concerns the sampling phase. There are two main questions involved in this first phase in the social sciences: First the analyst needs to decide what individuals, institutions, channels, receivers, etc. should be taken into account given the research purpose. The second relates to the decision on whether partial or full extraction methods should be used, for instance in the questionnaire preparation. To give an example: If one's research focuses on the right extremism in France as presented in the press, the decision could be to collect all news, reports and commentaries which appeared in newspapers in France, say in 1996, reporting on this matter, or in general all newspaper material of that year. A different example, leading to different sampling and collection criteria - plus different kinds of texts - is the analysis of free (open) responses to a survey concerning right extremism of a population sample of teenagers between 14 to 16 years old.

With sampling the social scientist is interested in the adequate coverage of ideas or opinions. Sampling decisions become complicated by the fact that in most - but not all - cases content analysts aim at inferences, descriptions or findings which are explicitly or implicitly relevant to a larger communication body or textual basis than the sample which has been analyzed.

There is a rich literature on sampling methods and types, (stratified sampling, one-stage sampling, etc.), efficiency of sampling methods, technical aspects of sampling from which the content analyst can draw advice. To go into the details of this phase would be a diversion from the purposes of this report. It needs to be borne in mind, however, that sampling is related to the purposes of analysis and that the decisions made at this phase are directly related to the analysis results. The reader is also referred to section 3.1. for sampling related issues in connection to exploiting text features.

4.2 Corpus construction: data conversion, indexing, tagging

The basis of text analysis, i.e. the text data or corpus, may consist of spoken texts, e.g. interviews, monologues or dialogues during psychotherapy sessions, advertisements, or it may consist of written texts, e.g. answers to a questionnaire, newspapers articles, letters, etc., or a combination of written and spoken text material. For computer-assisted analysis the text corpus needs to be in machine-readable form. Therefore, if the corpus at hand is not machine-readable already, transcription procedures including development of transcription schemes, scanning and proofreading techniques are applied.

Once the text data are electronic, text management is performed. This may vary depending on the purposes of analysis, available computer tools, technical support, and required level of detail. Typically, texts are indexed, incorporating 'text-source-specific' information: single text or a collection of texts, line, sentence, speaker, text source, date or time period, language etc. can all be indexing criteria. Such information can be used for the statistical analysis.

In computer-assisted text analysis the constructed corpora are rarely - if ever - encoded according to the ISO standard SGML, i.e. Standard Generalized Markup Language (see ISO, 1986 and Goldfarb, 1990 on SGML, but also Sperberg-McQueen & Burnard, 1994 on the guidelines for the encoding and interchange of machine-readable texts). SGML has gained a great acceptance in the last decade for encoding electronic texts as well as assisting in the interchange of electronic texts. Although a large amount of effort is invested into collecting, transcribing, proofreading, and preparing the textual sources for computer analysis, no effort is made in producing (or alternatively making use of already existing) SGML encoded texts, in order to profit from the advantages such

standard and general document mark up can offer not only for own analysis purposes, but also for re-usability, sharing and dissemination purposes. In fact, exchange, re-use and/or further exploration of a constructed corpus by another analyst or group of analysts does not appear to be a consideration. The possibility of exploiting the structural information of texts or the coding performed by another analyst by means of re-using and perhaps refining it, as well as the possibility to perform comparative studies based on the same material is not explored.

This raises the question about the necessity of a central organization, a kind of social sciences text data archive. At the moment although numerical data which are the results of sociological investigations and surveys are archived, administered and made available by such a central organization in Germany (Zentralarchiv), no analogous organization exists to collect, archive and administer textual data (questionnaires, texts, interviews, etc.) either on a national or an international level. The question which arises is whether the non-existence of such an organization is due to lack of interest in re-usability of textual material either for the purposes of reproducing existing studies, or as data for secondary analysis purposes. A further consideration to be made concerns the 'service' the social sciences would offer to other disciplinary fields which might make use of that material, such as pedagogy, psychology, linguistics to mention just a few.

4.3 Categorization scheme - dictionary

It is acknowledged that content analysis stands or fails depending on the quality of its categorization scheme (Berelson, 1952). It is to be expected that the validity of the analysis results depends on the validity of the categorization scheme used for analysis. It is not surprising, therefore, that the development of the scheme to be used for analysis is a task of primary importance. A categorization scheme forms the basis of the analysis of content. It contains the categories according to which portions of texts - typically words, but sometimes also phrases - are coded (tagged). One can view a scheme as a thematically relevant word list, and often word lists and categorization schemes are used to denote the same, namely a dictionary. The word or word list for each category of the scheme is the basis of a dictionary.

For the development of categorization schemes for content analysis two 'ideal' types may be distinguished:

- a set of relevant categories is developed, which is based on the researcher's understanding of the context of situation she is interested in analyzing. According to this scheme, portions of texts - typically words in computer-assisted analysis as opposed to words or phrases in conventional content analysis - are coded. In general, the development of a scheme for interpreting textual data relies on the analyst's interests, purposes of analysis, intuitions, expertise and experience. Note that whereas in conventional content analysis the scheme has to be constructed and finalized before the coding takes place, although some refinement and modification may follow a pre-test phase, computer-assisted text analysis enables a degree of circularity with regards to categorization scheme construction and analysis. Key Word In Context (KWIC) lists are used to check whether the words used for categorization are valid, unambiguous and exclusive of each other, or whether the categorization performed according to scheme is error-free and if not what kinds of errors occur, which can be probably corrected. Analysis examples making use of this approach provided in the General Inquirer book (Stone *et al.*, 1966), the General Inquirer being a program which supports this approach.

- In comparison with the first type, the second one is not theory-bound, but rather 'data-driven': the categorization scheme is constructed a posteriori, based on the content words of the textual material. More specifically, one attempts to generate from the available textual

data those categories which are more relevant or significant for the analysis. The relevance or significance is equal at a first stage to high frequency of occurrence of words and to clusters of words at a later stage; by applying particular statistical procedures, such as cluster or factor analysis for instance, from a(n arbitrary) number of the most frequent words or groups of 'synonyms' one generates relevant scheme categories for the analysis purposes. In that sense, the text corpus itself determines the categorization scheme to be used, based on the significance (indicated by frequency of occurrence totals) of content words.

Regarding the a-priori type, the dictionaries used may be specific in which case it is appropriate for examining a fairly narrow and well-defined issue, such as the Anxiety Theme Dictionary developed at Ulm University (Grünzig, 1983), or a dictionary may be general, i.e. it is used for examining general issues, such as the Harvard Third Psycho-sociological Dictionary (Stone *et al.*, 1966). Some well-known dictionaries used in computer-assisted content analysis are the Lasswell Value Dictionary (Lasswell & Namenwirth, 1968), the Harvard Third Psycho-sociological Dictionary, or the Regressive Imagery Dictionary (Martindale, 1978).

4.4 Coding

Next to categorization scheme development, coding is the other main process of content analysis: Based on a categorization scheme, words or phrases are given a code. As already mentioned (section 2.1), depending on the mode of analysis, coding is either an intellectual (manual) task or partly manual and partly computer-supported, whereby human coders are provided with on-line support for code assignment, or coding is performed fully automatically. For the latter, word ambiguity, metaphor usage, pronominal reference and its resolution, etc. are all sources of complexity.

Ambiguity relates to polysemy in language. A word or a phrase has more than one meanings depending on context and usage. For a computer program to decide what is the right meaning and subsequently assign a text unit to the appropriate code, it needs often to make decisions based on grammatical features, i.e. what part of speech category a word belongs to, as well as the context of communication situation, i.e. who the speaker/writer is, what is the topic, what semantic domain the topic belongs to, what the mode of communication is, etc.

An important point to make is that coding does not mean that every single word of the text corpus must be coded; rather only those instances which are relevant to the research goals and relate to the categorization scheme used. Considering this, an additional source of complexity in computer-assisted analysis with fully automatic coding is to determine what has *not* be coded, but *should have been* coded. Checking coding and evaluating its accuracy does not mean discovering only the 'miscodes', but also discovering those instances which one would expect to be coded as relevant, but which for some reason have not been coded. Corpus exploration routines are required for this purpose as well as means for defining search pattern rules (see Alexa & Rostek, 1997) where hypotheses are tested about sequences of words or phrases and how - or if at all - they have been tagged.

In contrast to the above, another necessary check for the accuracy of performed coding relates to filtering out those instances which have been coded, but in fact should not have been. KWIC views of the coded texts as well as means for checking the distribution of the immediate context of the coded occurrences are kinds of support for coding in content analysis.

4.5 Statistical analysis

In computer-assisted analysis statistical analysis follows coding. Coding has essentially resulted in reducing the textual data to a meaningful and small amount of numerical values, which can be further statistically processed.

The numerical values generated by coding and those variables related to the nature of the corpus, i.e. texts, speakers, age of speakers, etc., are passed on to statistical packages for statistical analysis: factor analysis, multidimensional scaling, cluster analysis. The analyst may return to the textual data - as an iterative process - if the validation, modification and refinement of the scheme and the dictionary (or dictionaries) used is required. The basic means for exploring the textual data further is that of obtaining and investigating KWIC displays. What needs to be born in mind is that analysis in the sense of interpretation or indication of what the codings mean, always means that some sort of statistical analysis is performed. Interpretation is tightly related to interpreting the numerical data by applying various statistics.

In contrast to the manual (coder-based) analysis, it is not unusual in computer-assisted text analysis that the analyst returns to the coded text corpus either in order to perform a more specific investigation or to refine/modify her coding. This return is based on the analysis results of the first 'circle' of analysis. Given the high costs in both and time and effort, the return to the original data for further analysis or correction in manual analysis is very unlikely once the textual data have been coded and prepared for statistical analysis.

4.6 Sharing: Dissemination and reusability of coded corpus, categorization scheme and dictionary

In general, dissemination of textual data, either raw or coded, does not appear to be a consideration of text analysis in the social sciences. Text material is collected and coding is performed, without application of such means which can support or ensure the reusability of this corpus by other researchers or for other analytical purposes in the future. The sharing of textual resources is not discussed in the relevant literature or in the reported specific content analysis practice.

As a consequence, it is almost impossible to profit from the effort invested by other researchers for analyzing very similar or identical textual data, as well as to share already performed, or proof-read, coding in order to perform further, perhaps secondary, analysis. Moreover, the lack of provision for reusability of textual data results in difficulty - if not impossibility - to reproduce analysis and check results, or present comparatively different analyses according to different methodologies but based on the same corpus. Additionally, as already mentioned in section 4.2 above, in computer-assisted text analysis projects in the social sciences no effort is made in producing (or alternatively making use of already existing) SGML encoded texts, in order to profit from the advantages such standard and general document mark up can offer not only for own analysis purposes, but also for re-usability, sharing and dissemination purposes.

4.7 Reliability and validity

Concerns about both the reliability, i.e. the dependability, consistency and accuracy of results, and the testing of the validity, i.e. the degree to which analysis has measured what it is intended to measure of the performed coding, are unavoidable for text analysis. We shall not go into the details of both theoretical considerations and their application regarding the reliability and validity phases, as this is outside the scope of this report. We shall consider only some aspects of both reliability and validity, which are directly related to computer-assisted analysis.

In manual coding, no matter how explicit and systematic the coding rules are, it is almost unavoidable, that different coders will make different judgments. The requirement for explicit, unambiguous, precise categorization which is not subject to variant, individual (human coder) interpretation is crucial when coding texts. Using automated procedures for that purpose bear clear advantages over the manual coding. Automatic coding is performed on the basis of explicitly formulated and unambiguous logical conditions; these are in a sense the externalization of the analyst's internal interpretative framework. In order for a program or a system to categorize and code automatically it has to have some 'knowledge' of the concepts and conditions which an analyst or a group of analysts have and use to understand, and thus interpret, a particular communication context. In other words, automatic coding depends on the formal expression and representation of the analyst's knowledge, in the form of a number of concepts and conditions for their application. The analyst herself is challenged to understand and describe the interpretation she is seeking for. As a consequence the application of these expressed formulations for coding text excludes inter-subject variance.

Furthermore, a number of studies, e.g. Carley (1988), Franzosi (1990b), Hohnloser *et al.* (1996), etc., have shown that on-line coding, that is, not automatic but rather manual coding done in a computerized environment, bears advantages with regards to the reliability (and validity) degrees over manual coding.

Validating the coding, and hence the interpretation performed, entails attempting to ascertain that the analyst has understood the meaning of the linguistic units (or characteristics) analyzed exactly as they were meant by those who produced these units. The different kinds of validity (for example, face or content validity, predicative validity, etc.) and validity tests (for example, expert validity, known groups validity, etc.) will not be described here, as this a topic in itself, which is outside the scope of this report. Nevertheless, a point to stress here concerns the advantages and the added-value of procedures for text (corpus) exploration in computer-assisted text analysis; the checking of the coding performed as a means for testing validity as well as checking the reliability of this scheme and the possibilities to correct, refine or develop it. The application of a categorization scheme is central in content analysis. Validation of a scheme very often requires means to work with the collected body of texts in an iterative manner. Iterative text analysis can be efficiently supported in a computer-assisted text analysis environment.

5 Methods of approach: Computer-assisted text analysis methods in social sciences

In the following sections we present a small number of methods of approach to computer-assisted text analysis. The reader should not expect a comprehensive list of all methods; the choice criteria are not motivated by the intention to achieve completeness, but rather to show the range of variation of the methods. More specifically, the aim of presenting some of these is

- on the one hand, to survey what is considered as standard for computer-assisted analysis in the social sciences (sections 5.1 and 5.2) in order to gain an understanding of the main processes involved, as well as of the changes and the developments that have taken place over time and,
- on the other hand, to survey some of the methods which have recently been developed and used, which are indicative of the variety of approaches and diversity of problems and aims in the practice in computer-assisted text analysis.

5.1 The 'a priori' type - General Inquirer

Probably the best- and most-known approach to computer-assisted content analysis in the social sciences is the approach developed by Philip Stone and his colleagues and realized in the General Inquirer program (Stone *et al*., 1966, Zuell *et al.*, 1989). Their approach is representative of the 'a priori' type (see sections 4 and 4.3) for the construction of a categorization scheme and dictionary.

According to the General Inquirer approach to computer-assisted content analysis, primary importance is given to the theory being investigated: the theory determines the overall research design, the categories and subsequently the rules to be used for coding, as well as the kinds of inferences that might be drawn from the analysis results. Therefore, the analyst defines what categories are to be used, and the categorization scheme is determined by the theory used.

In general, the content analysis procedure is characterized by the following two steps which interact with each other: first, the specification of content characteristics to be measured and, second, the application of rules for identifying and recording the characteristics when they occur in the data. The content characteristics can be measured once they have been coded according to the dictionary used: a dictionary depicts the collection of content analysis categories the analyst has been defined. To each category, called 'tag' once it has been assigned, belongs at least one word. The computer tags the text by using those categories. All criteria for coding the text data are explicit and predetermined.

For the purposes of reducing the over-all size of the dictionary the following strategy is used: The computer first looks up a word in the dictionary exactly as it appears in the text, for example, the words *walked*, *walks*, or *walking* may each be assigned different categories. If the exact word, i.e. the grapheme, is not found in the dictionary, then the computer looks for regular prefixes and suffixes (e.g. *-ion*, *-e*, *-ing*, etc.), removes them if found and then the word, this time without the suffix, is looked up in the dictionary list.

Although the General Inquirer may be compared to a dictionary, as an idea the program is similar to a thesaurus; it is a set of concepts by which words are grouped as similar or dissimilar. Each concept is defined by a list of words or phrases and the list of concepts comprises the dimensions of the thesaurus.[5] All words listed in the dictionary that have been defined as representing a particular concept are assigned a number representing the concept. For example, the concept 'SELF' contains the words *I, me, mine, myself* and all take the number '01'.

The computer processes the data sentence by sentence and each word or phrase in the text is compared with each word or phrase in the dictionary. At the end of each sentence a string of numbers representing the concepts found by matching the text and dictionary words and phrases is assigned to the sentence. All words or phrases which occur in the text but not in the dictionary are stored in a separate list and can be viewed later.

The application of rules for identifying and recording the content characteristics when they occur in the data is significantly assisted by means of defining and performing 'co-occurrence tests', which search for patterns (sequences) of particular categories. The whole pattern rule can then be assigned a different category. For example, by means of the following test the matching occurrences to the specified pattern can be tagged according to the specified category:

```
IF OCCUR (77) AND (OCCUR (63) OR OCCUR (84)) $ PUT (90) $
```

[5] Note that the method makes no use of a general thesaurus of language - such as the Roget thesaurus or the WordNet semantic net. Only those concepts of interest to the analyst are used for coding.

which means "If tag combination 77 and 63 occurs or if tag combination 77 and 84 occurs, then add tag 90 to the list."

The General Inquirer approach appears to work well for such analysis tasks relating to theme analysis, but has been criticized (Carley, 1988, Franzosi, 1990b, among others) as not sufficient for text analysis tasks where the identification and coding of relations between different themes or categories is required, or for performing comparisons between texts where one is interested in finding out how concepts are related and not just how they co-vary across texts[6].

5.2 The 'a posteriori' type - Iker & Harway

As introduced in section 4.3, Iker & Harway (1969) developed a different approach to the one represented by the General Inquirer. Both Iker and Harway, psychologists, have concentrated on the analysis of psychotherapy interviews (communication between a psychotherapist and a patient). Primary motivation behind their methodology is the attempt to infer what a particular body of textual data is about, without having to supply a priori categorizations within which the text data are classified, which is the General Inquirer methodology.

Their technique concerns the empirical generation of concepts (themes) and not the coding of textual data. Important for their analysis is to investigate *"the relationship between the substance and the structure of the oral communications among psychotherapist and patient, their change with time and with progress of treatment."*(Iker & Harway, 1969, p. 381).

Given the above point of departure for their analysis, their technique, as supported by the WORDS system (Iker & Klein, 1974), aims at obtaining factors which correspond with or indicate the major content themes of the text data and involves the following sequential process:

1. Divide an input document into "segments", for instance page, paragraph, (arbitrary) equal numbers of words, units of interview time.
2. Remove all articles, prepositions, conjunctions and general function words.
3. Reduce all remaining words to their root form, i.e. lemmatization.
4. Group together words which have the same basic meaning and their usage is the same (synonymization process).
5. Generate a list of the remaining word types, sorted by frequency of occurrence and beginning with the highest frequency word; down-count the list to reach 215 different types[7].
6. Compute an intercorrelation matrix on this subset of 215 word types.
7. Factor analyze the intercorrelation matrix and
8. rotate it to simple structure against a varimax criterion.

For the intercorrelation matrix the correlation coefficient is used as a measure of the strength of association between any word pair. The values the correlation coefficient can take range from +1.00 through 0.00 to -1.00: if the occurrences of two words are independent of each other, then they are unrelated and their correlation is approximately zero. If two words relate in that the presence of one is associated with the absence of the other and vice versa the result is a high

[6] For some more points of criticism of the presented approach see also section 5.8.

[7] The total of 215 types is dictated by the system capabilities of that time. As reported in Iker and Harway (1969) "the maximum matrix with which we can work is 215 variables" (p. 384).

negative correlation, whereas if two words relate in that they tend to occur together and not to occur together the result is a high positive correlation.

The fourth step of the above process, namely synonymization, makes obvious allowances for subjective judgments, which is what this approach attempts to eliminate. The analyst is required to make subjective judgments - which are potentially unreliable - and this fact demands to a certain extent the same effort from the analysts as the a priori categorization approach to content analysis. As a solution to this problem comes what Iker & Harway (1969) euphemistically called the 'Untouched from Human Hands (UHH)' approach.

According to UHH, synonimization plays a minimal role; ad hoc decisions deriving from the analyst's own inspection of the data are avoided and instead a generalized set of rules is applied: During the pre-factoring phase a large number of words are deleted based on

- what part of speech category they belong too, e.g. articles, prepositions, etc. are removed,
- whether their content is insignificant, e.g. *sort*, *still*, *be* are examples given in Iker & Harway (1969),
- certain combinations of words occurring with a specific part of speech category, e.g. *kind* occurring as noun is deleted whereas as adjective is retained, and finally, the application of what the authors *call "low level of predetermined synonymization"* (Iker & Harway, 1969, p. 386), according to which generic words are created to subsume a set of other related high frequency words, e.g. 'NO' is held as the generic word for the occurrences of *neither*, *never*, *nobody*, etc.

Iker (1974) presents a technique developed for the selection of the subset of 215 word types designed to maximize the associational patterns among words, called Sum and Evaluate the Largest Exponentiated Correlation Terms (SELECT). According to this technique, all words remaining after the removal of function words and the lemmatization step which have a frequency of equal or higher than ten are retained for screening. During screening those correlations which they represent chance events and which should not be considered are removed after application of a significance test.

It is stated in Iker (1974) that when working with a large matrix *"the problem in using the size of a correlation sum as the criterion for selection is the potential "washout" of a (relatively) few large correlations by a host of smaller ones. A correlation of .900 can be overshadowed by three of .31, or by five of .181, or by 20 of .046. With a large number of words (the situation for which SELECT is intended), say 500, each word is correlated against 499 others. The large majority of these correlations are low and there is, consequently, a substantial pool of very small correlations always available to wash out a few large ones on a given word. Interacting with this pool of small coefficients is the fact that a word with a few very high correlations inevitably shows more very low correlations than does a word with no very high relationships."* (p. 315). The solution to this problem is to increase the size of large correlations at the expense of smaller ones. It is demonstrated in Iker (1974) that the technique to best achieve this, i.e. coefficient powering, is by fifth power exponentiation. By means of this technique SELECT raises the number of words with highest correlations and enables selection of a word list (a word subset) with higher correlations than those produced by selection based on frequency information.

The methodological advantages of the Iker & Harway approach are clear: attempt to reduce, or avoid, intuition as much as possible by extracting empirically the thematically relevant categories for analysis from the text data themselves as well as attempt to reduce the amount of time consumed in order to find meaningful concepts and their language realizations.

On the positive side, if one's analysis involves a very large body of data, whereby the extraction of the general themes or topics of the texts is the primary goal of analysis, then this

method of approach may prove advantageous. Note, however, that it says nothing about the connections or relations that may exist between the identified themes and, therefore, is not productive for such analyses where the relations between different themes need also to be identified.

The main assumption on which the Iker & Harway method of approach, as well as the WORDS system supporting this approach, rests, namely that *"there exists sufficient meaning within the word and within the temporal associations among and between words to allow the elicitation of major content materials and categories"*(Iker & Harway, 1969, p. 381), essentially claims *a single* word form denotes meaning but excludes that a sequence of words can denote a single meaning or that a single meaning is denoted by a word combination (i.e. a lexicalized word sequence). In the pre-factoring phase articles, prepositions, conjunctions and general function words are removed: this, however, can result in loss of meaning (and hence content) due to the fact that some part of their realization expressions is removed. To give a simple example, if one removes the preposition *up* then there is no way detect usage of another meaning as denoted by *make up*. It is well-known that one of the characteristics of language is its combinatorial capacity: words are not only used on their own, but they form groups with other words to mean something different.

The dependence on single graphemes may cause significant loss of information and in that way not be beneficial for the purposes of this approach, as one may overlook, or modify by overlooking parts of, what the data are about. The question the analyst may pose regarding this rather 'materialistic' view towards a body of texts is: Is the main focus of my analysis the language used and the meanings expressed in the collected texts, or do I want to perform a dictionary analysis, which comprises only a part of language used in the collected texts?

An additional remark concerning the usage of word stop lists is that one should always be aware of the implicit assumption made when 'non-content' words are removed that content can be assessed in isolation of form.

Undoubtedly, the attempt to generate empirically the set of themes or, also, a set of categories to which portions of a text may be allocated is advantageous and by no means criticized here. What is, however, criticized is how this attempt is realized and presents some of the risks entailed for not achieving the goal it set out to achieve. Enhancing the presented technique would require means to check how words are used in context, compare co-occurrence contexts and indicate presence or absence of identical 'meanings' *before* the lemmatization phase. To this end one requires such co-occurrence check operations, which can compute and extract the local contexts of the words and detect whether they are identical or not, examining, thus, the specific language usage of a body of texts.

5.3 Concept mapping

Miller & Riechert (1994) describe a method for content analysis, namely *concept mapping*. By means of creating 'visual maps' of the most dominant themes for a text data collection, Miller & Riechert (1994) attempt on the one hand to circumvent the problem of subjective reading and interpretation and, on the other hand, to speed up the analysis process by generating automatically those key terms which are dominant in the text collection.

Miller & Riechert (1994) claim that the concept mapping analysis methodology is designed to identify important issues and indicate their relationship to one another. Their textual analysis interest operates in the media analyis domain, where a large body of texts is usually analyzed and compared.

They illustrate their methodology with an example concerning the amount of media reporting as well as the emphases in the media coverage on the topic of risk of pesticide usage.

After the sampling and data collection steps and by means of using the VBPro programs to obtain a frequency list of all words of the collected set of documents, concept mapping applies a computerized procedure which chooses terms on the basis of their *"mathematical value"* rather than their subjective meaning. The basic assumption made is that a word is indicative of the theme or topic of an article to the degree that it has a *relatively* high frequency in that article. The VBPro programs automate this procedure via a Chi-square statistic: the program computes an expected value for the occurrences of a word in an article based on the percentage frequency of occurrence of that word in the combined set of all articles. This expected value is then compared with the actual number of occurrences in the article. These values are used in the usual Chi-square formula, that is, the difference of the observed and expected values squared, divided by the expected value, and summed across all cases.

Calculations of these values for the analysis purposes reported in Miller & Riechert (1994) are performed for the 1,024 highest frequency terms.[8] Terms with high square values are examined and those which are function words, such as articles, prepositions, conjunctions and auxiliary verbs which are assumed to carry no substantive meaning although they frequently occur in texts, are removed. This procedure generated a total of 121 words with the highest chi-square rank which are used for analysis.

Each word is tagged for its frequency of occurrence in each article (using the VBPro programs). Furthermore, each article is coded for the specific magazine it was published. The result of this coding is a data matrix of 161 rows (one for each article) and 125 columns (1 for each word and one for each publication source, i.e. each title).

The coded data set is then the input to the concept mapping procedure: this calculates a matrix cosine coefficients that is indicative of the degree of co-occurrence of the words comprising the data set. The largest three vectors for each term of the set are then extracted from the cosine coefficient matrix. The values of the first vector depend primarily on the frequency and number of co-occurrences of each term. These values are interpreted as the prominence of that term. The second and third vectors are interpreted as the dimensions to project the words into a two-dimensional space.

In order to make the results easy to understand and avoid having them appear cluttered on a two-dimensional map, the output of the mapping is cluster analyzed in order to determine the kinds of themes they indicate. The generated map is based on the average positions of the clusters and their relative size.

Miller & Riechert (1994) conclude that the concept mapping methodology is beneficial for both their specific analysis as well as in general, due to the fact that it readily identifies dominant themes, their relative importance and their relationship to one another in a format that is quickly interpretable.

Note that the concept mapping approach is strictly word-based; in fact it appears that it is strictly grapheme-based. One is tempted to detect a certain 'insensitivity' towards language: no lemmatization procedure is reported prior to generating the list with the most frequent terms. Since there is no means for grouping content words referring to the same semantic 'sphere' together

[8] No details are provided about the particular decision on the total of terms to be considered. Also, no frequency total is provided to explain what the highest frequency is.

(farmers, farming, agriculture, etc.), there can be information loss. Although, in principle, the attempt to extract empirically the topics or themes or categories for analysis bears similarities to the Iker & Harway (1969) approach (section 5.2), it differs from it in that no synonymization procedure is performed, i.e. no grouping of semantically similar words is attempted.

Moreover, no tests are considered to check polysemy of the terms included in the data set: it is well-known that language is ambiguous: if polysemous usage or, in general, language ambiguities are not recognized this can easily produce misleading results. No step is reported in the paper on how to test this. Design and inclusion of additional operations to check whether the local or immediate contexts in which each the term of the data set is used in the texts under consideration are identical would enhance this method. This would take specific language usage into account and benefit from text exploration procedures.

Additional to being word-based, concept mapping is also part of speech based. Prepositions, auxiliaries, articles and conjunctions are removed as function words. Similar to Iker & Harway (1969) the criticism here is that by removing this, without having prior used such operations to test whether some a group of words (two or more) are used as a single semantic unit, as collocations or idiomatic phrases. The danger here is that by counting all occurrences of a given word as one meaning, some meanings are ignored and some meanings are given more importance than they should have to. Design and implementation of additional operations to check this aspect would enhance the method.

5.4 Minnesota Contextual Content Analysis

McTavish & Pirro (1990) and McTavish *et al.* (1995) have developed a computer-assisted content analysis method for the measurement of the social distance between positions (statuses) in an organization. Starting from the fact there are evident language-specific, stylistic differences in the everday interaction of say managers and employees or doctors and patients in organizations such as firms or hospitals, McTavish and his colleagues suggest to measure social distance between statuses in organizations by means of analyzing verbatim transcripts of open-ended conversations of persons of a number of social positions within a particular type of a social institution. Analysis involves the coding of social perspectives. The differences between scores in terms of social perspectives is a function of a social distance between different statuses. Each verbatim transcript is scored for social distance using the computer content analysis program Minnesota Contextual Content analysis (MCCA), which indicates the similarity to themes characteristic of broader institutional perspectives.

Distinguishing feature of the MCCA approach to computer-assisted content analysis for the coding of textual data is that contextual information is taken into account. The MCCA computer program entails a dictionary whose construction is typical of the a priori type of categorization scheme/dictionary construction (see sections 4.3. and 5.1). However, unlike the 'standard' content analysis methodology, the dictionary contains not only conceptual, but also contextual information since the dictionary is augmented with four different contextual dimensions (vectors), namely *traditional*, *practical*, *emotional* and *analytic*. Each contextual dimension incorporates a general idea of social activity and represents a different framework within which specific concepts can emerge. McTavish & Pirro (1990) find that these dimensions *"satisfy several criteria we consider important for any set of contextual" markers" used in social science investigation".*

Each context comprises a set of categories (117 in total, with one category being the leftover list of words which have not been categorized), e.g. *emotional* incorporates the categories 'Happy', 'Pleasure', 'Expression Area' and 'Self-other'; each category in turn contains typical words or phrases, e.g. *gladness* and *refreshment* are typical words for the category 'Pleasure'. Ambiguous

words are contextually disambiguated to determine the most appropriate category, and each category receives some differential weight reflective of its usage in the different contexts.[9]

Each word in a text is matched against the concept categories of the dictionary, keeping a running tally of usage, concept by concept. Conceptual categories tallies are percentaged for each text by the total words in the text. This total is subtracted from an expected score obtained from a norm to yield an emphasis score (E-score) for each of the category concepts; these scores are the basis for conceptual analysis. McTavish & Pirro (1990) give an example of how this is calculated for the sentence 'Work like mine keeps me from doing my best':

"the idea of 'self' appears three times (e.g. mine, me, my) out of nine words or p(i,k) = .333.[10] If the expected occurrence of this category is P(i,k) = .045 and the standard deviation of this category's usage across contexts is S(i) = .028, then the E-score can be calculated as follows: "Self" E-score = 0.333 - 0.045 / 0.028 = +10.29. This suggests that the idea category "self" occurs more frequently than one would expect if the nine words reflected the usual English conversation."[11]

Contextual scores (C-scores) are also calculated: As each word is identified and classified into a conceptual category, four cumulative contextual scores are each updated, whereby the updating procedure uses weights which reflect the **relative use** of each category in the four general social contexts. Accumulated C-scores over a text are standardized. In that way distances between texts according to the space of the four context categories can be computed and used to express the proximity of texts to each other. By means of cluster analysis one can display the structure of the proximity matrix.

McTavish & Pirro (1990) acknowledge that further theoretical and quantitative work is needed on linkages between conceptual definitions of key social science variables and patterns of word usage, as well as on expectations for comparative word patterns across cultures, societies, institutions, organizations and historic time.

We may distinguish two major features of the MCCA approach: First, inclusion of contextual information, although this is by no means (language) theory grounded; the four contextual dimensions are drawn based on personal judgments and experience rather than being based on a 'theory of context'. In fact, it would be interesting to investigate whether and to what extent applying a general theory of context or using lexical semantics would improve the method. The second feature concerns the consideration of expected frequency of words in general language; this emphasizes the necessity of taking into account the different contexts of situations. Furthermore, it acknowledges and exploits the fact that language use differs according to situations, which may have as a consequence that average frequency of occurrence of some words may be loaded for certain contexts when matched against some neutral expected standard.

A recent development (see McTavish *et al.*, 1995, Litkowski, 1997) within the MCCA approach is the correspondence of the MCCA dictionary categories to the WordNet™ synsets. WordNet (Miller *et al.*, 1993) is a semantic network of about a hundred thousand words, which are grouped in synsets, that is sets of synonyms. The synsets are connected with one another by a

---

[9] No further information is provided in the relevant literature about how the disambiguation process operates, i.e. manual or automatic.

[10] *p(i,k)* is the observed proportion of text in conceptual category *i* for text *k*.

[11] As a general norm or expectation for language use in a broad American English conext, McTavish & Pirro (1990) use probabilities associated with word use. These probabilities are based on the counts and percentages for the Brown corpus of American English as provided by Kucera, H. & Francis, W.N. (1967): *Computerized dictionary of present-day American English*. Brown University Press, Providence, RI.

number of semantic relations, such as antonymy, synonymy, part-of, hyponymy, etc. All WordNet nouns and verbs are hierarchically organized into about 150 semantic categories. McTavish *et al.* (1995) have shown that MCCA dictionary categories are consistent with WordNet synsets, and that they appear in fact as supercategories of the WordNet synsets. The usage of the WordNet data is shown to be advantageous in two main ways:

- it allows the analyst to move from subjective conceptual or contextual groups to more general semantic groups which reflect fine-grained meanings inherent in particular words and
- it enables the further refinement of the idea categories into semantic components: it is possible to extend the amount of words that might be associated to an idea category, mainly by identifying and including the narrower terms (hyponyms).

Both of these may result in explicit and transparent analysis.

Litkowski (1997) maintains that tagging with MCCA categories as well as the WordNet synsets, although valuable, is insufficient. MCCA uses multidimensional scaling (MDS) in order to produce a map when given a matrix of distances. MDS analysis of C-scores provides a first characterization of texts and identification of the contextual focus: the analysts may thus identify concepts and themes that are different and similar in the transcripts analyzed. MDS analysis of the E-score vectors identifies the primary concepts that differentiate the transcripts analyzed. The graphical output of analysis is examined in order to label those points with dominant MCCA categories. The 'meaning' of the MDS graph is then described in terms of category and word emphases, and these are the results used on reporting on the content analysis of texts. Litkowski (1997) points that this is exactly where the weakness of MCCA categories become obvious, since the interpretation of the MDS analysis output is subjective and based only on the identification of particular sets of words that distinguish the concepts in each text. If additional lexical semantic information were used to enrich the MCCA categories then this is expected to result in achieving greater sensitivity for the characterization of concepts and themes as well as in analysis which would be performed based on more rigorously defined principles.

5.5 From part of speech tags, to syntax to semantics

A different approach from the ones presented so far, which incorporates syntactic and semantic knowledge for the analysis of a corpus of free responses to open-ended questions in social survey research, is suggested by Nazarenko *et al.* (1995). Their research applies computational linguistics techniques for tagging syntactically and semantically a corpus of answers[12]. This annotated corpus is then statistically analyzed. The difference in their methodology then is that statistical methods are used for the analysis of linguistic information as opposed to the standard statistical analysis of words (what they call "graphic forms") or theory-based scheme categories.

Specifically, the following analysis procedure is followed: The first step is to partition the corpus comprising answers to open questions according to criteria applying to interviewees, namely age (under thirty years old, between thirty and fifty years old and above fifty years old) and education level (high school drop-outs, high-school graduates, academic education) This partitioning is necessary, since *"statistical methods cannot apply to such small fragments of text"* (Nazarenko *et al.*, 1995, p. 30). In order to characterize a given partition they attempt to capture the forms which have a higher frequency in the partition than in the rest of a corpus. Nazarenko *et al.*

[12] It is worthwhile to note that the corpus used for their analysis is in fact re-used: Nazarenko *et al.* (1995) are interested in comparing the results of their approach to those of Lebart & Salem (1994) and therefore, they have worked with the same corpus. This relates to the points raised in section 4.2. with regards to text data availability.

(1995) use the specificity method of Lafon (1984) to identify these forms. This method calculates the specificity *Sp* of a form *x* being an approximation of the probability for this form to have the frequency *f* in a part, given its frequency *F* in the whole corpus. The specificity method can apply to single graphemes or sequences of graphemes, containing no delimiters. In that way an *Sp* value is obtained for each corpus partition for each form or idiomatic expression.

Next, the corpus of answers is tagged with part of speech information by means of using a probabilistic part of speech tagger. The resulting tagged corpus is manually corrected. Each word form in the tagged corpus is associated with its lemma and morpho-syntactic tag. Therefore, each open response is expressed as a sequence of morpho-syntactic tags which are the statistical analysis units.

Nazarenko *et al.* (1995) are interested analyzing statistically tag sequences; statistical analysis, however, is difficult to interpret due to the nature of the tagset, being designed specifically for parsing purposes. The original tagset is, thus, simplified by eliminating the morphological features, which are not relevant for the analysis, masking them in the labels of the tagged corpus. A further modification is performed, whereby adjectives are subcategorized according to their syntactic properties. The resulting corpus after the above modifying operations have been performed contains a substantial amount of specificity in terms of morpho-syntactic tags and it provides an initial indication of the important syntactic features of the corpus.

Next step in the analysis process is segment analysis, which considers not atomized units as the previous step, but rather patterns of categories of the tagset used. The specificity method is applied to tags and tag segments which helps to identify the syntactic patterns which are characteristic of the corpus. This aims at uncovering fine characteristics which are not easy - if possible - to detect when analysis is based on graphic forms only.

The semantic tagging of the corpus follows the segment analysis phase. Semantic tagging is divided between terminological tagging and thematic links encoding. The first aims at connecting nouns to their hypernyms or synonyms. Only nouns were terminologically tagged. Similar to morpho-syntactic tagging, the purpose of semantic tagging is to enable the statistical analysis of semantic entities and not, as in the typical content analysis, to interpret the semantic content of graphic units.

The groups of synonyms and hyponyms are small, independent and shallow noun hierarchies which are encoded in a KL-ONE[13] like network, where the word lemmas are the nodes, synonyms are defined as variants of a given node and hyponyms are represented by means of subsumption, and a semantic tag is defined as a node property. The tagging groups a hypernym and its hyponyms under a single tag and these tags substitute then the word lemmas in a text, i.e. an open response.

The second part of semantic tagging comprises the manual coding of the texts for thematic information. In open response surveys it is important to detect those themes which are prominent in the answers, since these themes are the natural candidates for items of a closed response survey. Thematic information means the grouping of all graphic forms which introduce the same theme under a single semantic tag. The building of thematic relations is also guided by the hierarchies already generated by means of terminological tagging and therefore if a hypernym is assigned to a particular theme then its hyponyms are assigned to the same theme and the thematic links of every hyponym can be automatically derived. Thematic links are included in the noun hierarchy as new node properties and a second semantic rule groups under the same semantic tag all word lemmas that have been assigned to the particular thematic property.

Nazarenko *et al.* (1995) use both manual and automatic coding: thematic links as well as semantic features are coded manually whereas part of speech tags are not. Beside the well-known risk for inconsistent coding as well as the large amount of time and effort required for manual tagging, it is nevertheless worthwhile to note what advantages there exist for analysis when higher-level linguistic information is incorporated. Nazarenko *et al.* (1995) conclude that syntactic information *"helps to bring out the specific patterns of each category of interviewees"* and yields such results *"that could not be easily derived from statistics on the graphic forms or even on segments"* (p. 38). Although they acknowledge the difficulty involved in automating semantic tagging, they conclude that (linguistic) semantic information *"advantageously describes subjective information in a systematic, stable and explicit way"* (Nazarenko *et al.*, 1995, p. 38).

The generation of word groups according to hyponymy and synonymy relations and their network representation is a convenient and advantageous way to move towards generality and at the same time incorporate conceptual information; thus, the underlying meaning can be included in the analysis and single dependence on word forms can be sidestepped. This methodological aspect is promising and it will be interesting to see it applied in other content analysis contexts.

It is not very clear what contribution the thematic links make. On the one hand, it is shown to agree with the statistical analysis of graphic forms. On the other hand, the thematic links are intuitive, not based on a specific context-dependent theoretical framework. Perhaps the exploitation of general conceptual organizations (ontologies) rather than the construction of thematic links by Nazarenko *et al.* (1995) could provide a richer and objective basis for analysis, plus reduce the effort of devising and drawing the thematic links.

5.6 Automatic content analysis of spoken discourse

Wilson & Rayson (1993) report on their methodology for automatic content analysis, which is the result of a collaboration between a market research firm and a university center for corpus linguistics with the aim to develop an automatic 'content analyzer'. Noteworthy in their approach is the mixture of linguistic and specific semantic domain tagging.

The methodology used is as follows: interview spoken data are transcribed into machine readable form and each text is identified by mark up in SGML and divided according to the questions in the interview as well as features of the respondent, e.g. age, sex, social group. Following this, a manual, minimal linguistic mark up is performed for resolving reference relations, e.g. the antecedents of the pronouns 'it' and 'they' are explicitly provided (encoded in SGML). The textual data are then run through a spell checker, whereby misspelled words are not matched in the system's lexicon.

Automatic part of speech tagging is subsequently performed by using the CLAWS system (Garside *et al.*, 1987). The part of speech coded output forms then the input to the semantic tagging procedure. *Each* lexical item[14] in the text is tagged with a semantic tag. These semantic tags are loosely based (although now modified) on Tom McArthur's Longman of Contemporary English (McArthur, 1981) on the ground that this *"appeared to offer the most appropriate thesaurus type classification of word senses for this kind of analysis."* (Wilson & Rayson, 1993). This phase of semantic tagging uses a lexicon which consists of lexical items each of which has one syntactic tag and one or more semantic tags assigned to it. Structurally ambiguous words, e.g. 'offer' as a verb

[13] KL-ONE is a knowledge representation language.

[14] Closed class words such as prepositions, conjunctions, etc. and proper nouns are assigned a tag which indicates a grammatical 'bin' category. Apart from proper nouns the other words are not taken into account for the final statistical analysis.

and as a noun, are duplicated with a separate entry for each part of speech category. For the semantic tagging of the textual data then both word form (or stem) and part of speech category are used.

Some automatic disambiguation is performed for those words which have been assigned more than one semantic tag. This is done by means of weighting by domain of discourse and augmented with a rank ordering of sense likelihood for the language as a whole, but in general the resulting disambiguation is not satisfactory enough. Manual post-editing takes place for the remaining ambiguously tagged as well as unmatched items.

A significant aspect of the approach reported by Wilson & Rayson (1993) it the exploitation of a sample corpus of market research interviews for the development of specific rules for linking negative words to the items they negate, linking modifiers to adjectives or adjectives to the nouns they modify. For the rule development they extract recurrent sequences of part of speech tags and check the incidence of the relevant links. These links form chains of relationships linking, for instance, the modifiers with the relevant adjectives, so that when frequencies of a particular tag are provided, frequency of a word together with a particular modifier is also provided so that *"the qualitative aspect is not lost in trying to lump boosters together in numerical value"* (Wilson & Rayson, 1993).

Wilson & Rayson's approach stands in contrast to keyword analysis methodologies, where mere frequency counts on the content categories for the texts being studied are provided, which, however are not informative about the which items particular adjectives refer to, or whether and how adjectives are modified by intensifiers or downtoners, and whether or not assertions are negated. In addition, it attempts to use linguistic information and and apply natural language processing techniques in order to extract 'relational' information which holds between different grammatical categories.

5.7 Map analysis

Map analysis (Carley 1993, Carley & Palmquist, 1992) is a class of methods which focus not only on analyzing textual data according to a set of categories or concepts - which is the typical in content analysis - but, also, on the relationships between the defined categories. In that sense, map analysis subsumes content analysis.

A software program that assists this type of analysis is the MECA toolkit. We shall not describe the program here, but rather focus on the new element map analysis brings in our analysis survey so far, namely the incorporation of relational information between concepts, both as definition as well as their frequency of occurrence.

Carley's reported analysis interests concern the determination of the similarity or the ways of difference between large numbers of texts across subject or across time, concentrating on comparing coded texts. Although word usage distribution across texts or text similarity in terms of the proportions of types of words can be addressed by taking a content analytic approach, Carley (1988) argues that *"the focus on concepts implicit to traditional content analysis often results in an overestimation of the similarity of texts because meaning is neglected."* (p. 79). Similar to the linguistics-based approaches (e.g. Roberts, 1989, Franzosi, 1987), Carley argues that it is necessary to preserve the semantic and syntactic structure of the text, and for that purpose it is necessary to examine and code also how concepts are related and not just how they covary across texts.

The basic goal of map analysis is, given a set of texts and a set of concepts, to determine for each text whether and which of these concepts occur, as well as the relationships between these

occurring concepts. Detecting similarities and differences in the content and structure of texts may be assisted by frequency of co-occurrence information and estimations and differences of the distribution of concepts and the relationships among them across texts.

Crucial in map analysis is what a concept and a relationship is: a concept can be a word, a phrase or a fact. Concepts can be organized into types and can be hierarchically organized in sub-concepts (narrower concepts) or super-concepts (broader concepts). A relationship is a tie or a connection between concepts and it can be a single word or a clause, e.g. *loves*, *less likely than*, etc. Relationships have "*strength"*, "*sign"*, "*direction"* (in Carley, 1993 or "*directionality"* in Carley, 1988) and "*meaning"*, and different relationships diffe according to them. Strength denotes the level of emphasis of a relationship; for example *loves* is stronger than *likes* or *good* is not as strong as *very good*. Sign denotes the difference between positive and negative relationships between concepts; *does not love* indicates a negative relationship. Direction denotes whether a relationship is unidirectional or bi-directional. Furthermore it indicates the direction of the relation between two concepts: in the phrase *Peter phoned you* the direction of the relation is from *Peter* to *you*, whereas in the phrase *You phoned Peter* the relation goes from *You* to *Peter*. Finally, two relationships vary in meaning if their respective verbal phrases denote different processes, e.g. possessive (has), emotional (love), etc. Two concepts and the relationship holding between them comprise a statement.

By means of the above, using map analytic techniques in text analysis results in a network of interrelated information with the concepts being the nodes of the network and the relationships the ties, with their particular 'values', being the ties between nodes. The amount of information to be recorded for each relationship is the researcher's choice. More specifically, the analyst can choose to simply record that a tie exists between two concepts, which implies that two relationships have the same strength, directionality, sign and meaning. Or, alternatively, the analyst may decide to record the differences in all or some of the relationship values for strength, sign, direction and meaning. Naturally, the more information to be recorded, the more time-consuming and complicated the coding task becomes. Nevertheless, preserving a large amount of information, clearly enables more detailed comparisons.

It was mentioned earlier that map analysis subsumes content analysis This is advantageous, since one can combine the techniques of both methods or move from map analysis to more traditional content analysis using the coded data obtained from map analysis. Moreover, the analyst can alternate between abstracting from texts by analyzing statistically the coded texts and remaining close to the textual data and examining what specific concepts occur in text and with what specific relations are they connected to each other.

Unlike computer-assisted content analysis with fully automatic coding mode, map analysis is harder to automate, since the coding of the relationships is harder to automate.

A significant feature of map analysis techniques is that they offer a solution for analyzing texts not only in terms of the concepts which they entail, the frequency of occurrence and the distribution of these concepts in text and their statistical interpretation, but also in terms of the connections there exist between these concepts. It is argued by Carley (1988) that map analysis *"is distinct from other approaches used to analyze text because it is based on a cognitively motivated knowledge representation scheme and a cognitive theory of knowledge, and because it extracts meaning as operational definition."*(p. 219). It would be worthwhile to investigate the possibility of using language-based ontologies or lexico-semantic networks for supporting the construction of the conceptual scheme and the relations holding between concepts which will be used for analysis as well as for partly automating the coding task.

5.8 Application of semantic text grammars

As the last method of approach we present Franzosi's (1989, 1990b, 1995, 1997) application of semantic text grammars as coding schemes. Although termed by Franzosi "computer-assisted", this approach can be categorized rather under manual content analysis (see Section 2.1). Neither automatic coding nor integration of manual and automatic coding procedures are used; rather by computer-assisted one should understand the use of a computational environment to record and support the coding performed by coders. Although this report aims at presenting computer-assisted analysis methodologies, we believe that it is of benefit to include in this presentation this particular method of approach for the following reasons: (i) it presents a different framework, namely semantic text grammars for coding text, whereby the connection between specific categories is also coded, (ii) it contrasts with the map analysis approach (section 5.7), in that the relations are linguistic and not cognitive and (iii) it exploits text type information for more efficient coding.

The formal application of semantic text grammars (for text grammars see van Dijk 1972) in the quantification of textual (narrative) data is proposed by Franzosi as a powerful tool for content analysis (Franzosi 1989, 1990a,1990b, 1995, 1997). Franzosi's text analysis interests lie in collective action research and his textual data are newspaper texts. In such research one is concerned with what kinds of actions are reported in the texts, and what type of actors perform these actions. To determine this, events and event characteristics must be extracted and analyzed. Analysis techniques which focus on keyword search or theme determination are not sufficient for his analysis purposes, since the collation of information into specific events is not possible. As exemplified in Franzosi (1990b) "*if the word "strike" appears in ten different input documents, how many strikes can one actually count?*" (pp. 230-231).

Some of the problems which 'traditional' coding schemes in content analysis pose and which necessitate a different approach are presented in Franzosi (1989):

- inability to explicitly code the connections between categories; for collective action research this means that it is difficult to relate specific actors to specific actions.
- Highly abstract nature of scheme categories; due to this fact, there is loss of both richness and specificity: concept or words denoting 'micro-actions' such as *cheer*, *picket*, *hand out leaflets* are all subsumed under the general (macro-action) category *strike*.
- Highly abstract coding schemes increase the possibility of the human coders to make theoretical judgments in interpreting the text and thus may increase the risk for subjective interpretation.
- Dependency of schemes categories and their level of aggregation on the specific theoretical concerns of the researcher who designed them.
- As a consequence of the dependency of the coding scheme to the specific researcher and analysis goals, the coding scheme is not reusable for secondary or further analysis in order to study those questions which result after the first analysis.

An analysis methodology is, therefore, required which not only codes explicitly the connections between categories, but comprises, also, a coding scheme which does not carry theoretical bias, is general enough to enable secondary analysis, and allows detailed categorization. For that purpose Franzosi proposes using a linguistics-based coding scheme. Given that Franzosi's main research concern lies in the kinds of actions specific actors perform, that is with events, one requires a "*methodology that allows us to extract **functional** categories (i.e. properties of events) from the whole texts (e.g., from a newspaper article or a police report) or from a series of texts on an event, rather*

*than **syntactical** categories from each sentence in the text. We want to read text selectively, picking up only certain thematic units and certain semantic structures and leaving the rest behind."* (Franzosi, 1989, p. 270).

Franzosi takes semantic text grammars as a means to represent functional information on collective events. The grammar is represented as a set of rewrite rules. An event is defined "*as a set of actions performed at a particular place and time by some actor(s) against or in favor of some other actor(s).*" (Franzosi, 1989, p. 276). The text grammar is based on a simple subject/action/object structure (with possible modifiers), called 'semantic triplet' (Franzosi, 1997, p. 136). The semantic triplets are the building blocks of the semantic text grammar which is used for collecting data for the collective action analysis. These semantic triplets can be further aggregated into higher level structures with events being one of them.

Events and the different characteristics of actors and actions are all formally represented in a series of rewrite rules. For example, the rewrite the rules for events and semantic triplet are:

<event> → [{<comment>}] {<semantic triplet>} {<document key>}

<semantic triplet> → [{<comment>}] {<subject>} {<action>} [{<object>}] {<document key>}

The first rule defines that the nonterminal symbol `<event>` can be rewritten as one or more sets of semantic triplets. The nonterminal symbol `<semantic triplet>` is the combination of one or more subjects, actions and objects[15]. Rules of this kind comprise the semantic text grammar which Franzosi uses data collection and the testing of specifcic hypotheses related to his research theme. The texts are encoded according to the grammar and prepared for statistical analysis using a relational database.[16]

A point which separates this approach from the others presented so far is that when using semantic texts grammars for coding and analysis, the particular type of discourse or the genre the texts belong to, is taken into account. This is in the form of the text grammar including different levels of aggregation of individual-level information or a different number of modifiers for each symbol of the triplets, or a different vocabulary. For example, in Franzosi's analysis of collective action as reported in newspaper texts the highest aggregation level is the campaign, but if one were interested in analyzing life histories the highest level of aggregation may be the life of an individual, with the childhood, adulthood, etc. phases comprising the lower levels.

Furthermore, it is suggested that textual organization information about a collected body of texts - in Franzosi's work the "deep schema of news report" (Franzosi, 1995, p. 158)[17] - can be used to increase the efficiency of analysis. Franzosi uses the deep schema of news reports suggested by van Dijk (1988) according to which a news report comprises both a summary and a story, the story comprising further a situation and comments. A situation comprises episode and background with the episode comprising both main events and consequences while the background comprises context, i.e. previous events, circumstances, and history. The efficiency of analysis may be increased (without loss of information) by not coding those parts of a report where information is duplicated, i.e. summary and background: "*Each new article in a set reports the newest development of*

---

[15] The nonterminal symbol object in the rewrite rule for semantic triplets is enclosed in square brackets, indicating in that way that it is optional.

[16] For data entering the MS-DOS program PC-ACE (Program for Computer-Assisted Coding of Events) is used (Franzosi, 1990b). PC-ACE combines a front-end with a relational database management system. A data matrix of occurrences of relations among actors, actions, etc. within each event type can be produced.

[17] For the notion of schema see van Dijk (1988).

an ongoing event, or a series of events. The longer an event lasts, the more likely that the "background" section of articles dealing with the event become increasingly repetitious. (...) If a researcher's interest lies in the historical unfolding of events, rather than in newspaper reporting practices, most of the information labeled as background, in van Dijk's schema should be skipped, provided it had already been coded from a previous article This would lead to great savings in coding time. The more information of an article is repetitious of previous articles, the more time will be saved." (Franzosi 1995, p. 159).

Evidently, for the particular type of research, analysis aims and domain of research reported in Franzosi (1989, 1990a, 1990b,1995 and 1997), using semantic text grammars provides a solution for coding and extracting relevant information for the particular analyses. In comparison to traditional coding schemes, even very sophisticated ones, a text grammar provides explicit and rich information about the connections between the categories, a requisite for various research topics and not only collective action research. The mode of coding of this approach is computer-supported: coding is performed on-line and the computer provides an environment for fast and reliable coding. In fact, Franzosi considers impossible to automate the suggested coding: "*only human beings can perform the parsing required by semantic grammars. The approach to linguistics-based content analysis described in the article is based on human input. The computer provides only an environment where coding is faster, easier and more reliable"* (Franzosi 1990b, p. 232).

One may nevertheless consider the possibility of providing some automatic support for such coding. Based on the provided grammar and on a pre-coded, 'test' text corpus, one could attempt to build in semi-automatic procedures, where the system makes a suggestion and the coder accepts or modifies it. To what degree this would be cost-effective would have to be investigated.

With regards to the particular words or phrases denoting types of action (as with the example of the word 'strike'), it would be interesting to see how far one could exploit lexical/semantic word nets containing relational information for the different types of actions, or a linguistics-based ontology for defining and categorizing different types of processes (verbs) and the different types of actors they involve in order to have a rich and general language framework for the coding (and in effect the acquisition of) events, as well as a basis to attempt to automate partly the laborious work of coding.

There is a degree of overlap of the particular output of Franzosi's analysis to that of information extraction systems (see MUC-3, MUC-5, Grishman & Sundheim, 1996), which are tested and evaluated for the analysis of and the extraction of facts from news wire messages. Perhaps a further point of investigation would be to see whether, and, if so, how informative for the specific analysis purposes resources and techniques employed for information extraction systems may be.

Finally, the suggested approach, where deep schema or textual structure information is incorporated, seems to concentrate fully on text corpora containing texts of a single text type. An interesting investigation, would be to test its transferability and the degree of its cost-effectiveness to such analysis where text material belonging to different text types is analyzed.

6 Conclusions

All the surveyed methods of approach require that the text analyst make choices that affect how the texts are interpreted and the potential results. As a consequence each approach presented has its advantages and disadvantages. Although each may be unique in its own aspects, each one of them may prove not to be appropriate for every text analysis task. Different research questions and application contexts require or motivate different analysis techniques. Whereas one method might work perfectly for a given research goal it might be poor for another. The development of the

methods of approach surveyed in this report is not independent from the research context analysis goals each analyst defined.

Furthermore, it is becoming apparent nowadays that with the current technological developments, the higher degree of hardware availability and the easy of access to mahine-readable text, computer-assisted text analysis is not restricted to only automatized or manual mode. As Stone (1997, p. 50) observes *"Technology no longer constrains researchers to one or another mode of analysis. Instead it is feasible to consider a platform of desktop computer text analysis capabilities, that allows the researchers to draw upon sequences of procedures that best suit their proclivities, to follow up on leads as they uncover textual patterns, and to check the validity of their discoveries."* Moreover, the qualitative-quantitative distinction is becoming less clear with the researchers having the means to complement each approach and profit from a combination of the advantages of both.

With the growing availability of information communicated electronically nowadays text analysis is becoming steadily important for discovering meaning. With computers becoming increasingly accessible and with the development and availability of more sophisticated software in the last decade the computer-assisted content analysis has been employed and been put under test as a text interpretation tool.

The methods of approach presented in this report show, however, that computer-assisted text analysis as a general tool needs to move on from simply analyzing surface forms as well as from neglecting or disregarding the connectivity of words toward incorporating meaning in order to be able to answer different and complicated questions about meaning, strengthen the inferential potential text analysis offers, as well as expand its scope with regards to different kinds investigations.

The 'standard' categorization scheme which is used to code relevant units of information consists of categories, each of which is assigned a name which is only heuristic in nature and has no meaning apart from the semantic interpretation the analyst has attached to it. Although the categories of a scheme may appear internally consistent in that there is some underlying similarity between the words belonging to a category, this is only subjective and partial to the language. Therefore, means to incorporate more general and formal conceptual categorizations are required. Furthermore, the relations existing between words are necessary for 'deeper' kinds of analysis and therefore syntactic parsers, semantic analyzers, ontologies, etc. are some of the tools or resources to incorporate.

By including contextual and semantic information or different kinds of tagging such as, for instance, part of speech or tagging of anaphoric references, recent computer-assisted content analysis projects demonstrate the necessity to have available and benefit from additional linguistic meta-information in order to be able to ask interesting or 'meaningful' questions about 'meaning'. Using linguistic information is not a new idea; we have discussed how content analysts have anticipated linguistics to assist in the content analysis task (section 3.2). What is new, however, is the actual testing and attempt to determine what kinds of information are possible or relevant and the methodology for exploiting these kinds.

As an intellectual exercise it may be interesting to theorize about a method and its potential merits, put to test it, put it into praxis and profit from its advantages we need the software to realize the methodology. The computational considerations concern the development of software to support analysis in a flexible and user- friendly manner, general enough to be suitable for the different types of computer-assisted text analysis, enabling sophisticated modeling of textual data as well as coding according to one categorization scheme or more than one. Moreover, such software needs to support the inclusion of linguistic information either by means of dynamically calling external (to the text

analysis application) programs or storing and flexibly manipulating such information in addition to the data. Furthermore, such computer-assisted text analysis software is needed which can support in an integrated and flexible manner both the circular process of quantitative analysis and qualitative interpretation as well as accommodate the different kinds of meta-information required for deeper analysis. This poses serious technical requirements for complex and nevertheless user-friendly systems, but also for data modeling.

Future developments to advance computer-assisted analysis encourage a cooperation between disciplinary fields, realized in terms of knowledge transfer and interchange as well as complementation given that different kinds of expertise in, for example, linguistics and language engineering, computer and information science, mathematics, knowledge engineering, the social sciences and the humanities, are required. The development of systems for computer-assisted text analysis needs to be a multi-disciplinary effort if it is not to either repeat old processes dressed up in new 'technological' attire or have to re-discover knowledge which is already there. Computer-assisted text analysis expands over a broad disciplinary scope touching upon diferernt application and research fields as well as research goals. Isolated research with regards to both methodology of analysis and building of the tools to support the analysis is both cost- and intellectually-ineffective.

Specifically, the following are some suggestions for expertise and techniques which a future computer-assisted analysis software may incorporate and support:

- incorporation of lexical-semantic information
- transfer of natural language processing techniques related to both thesaurus and dictionary modeling and construction
- incorporation of more general dictionaries or concept networks
- (multi-level) linguistic information, e.g. morphological, syntactic, semantic, contextual analysis
- manipulation, management and maintenance of categorization schemes
- different manipulations and views of codings
- data visualization techniques
- modeling of textual data.

Naturally, no analyst or social scientist employing computer-assisted text analysis can achieve that on her own. However, the odds that social scientists engaged in fruitful dialogue with information scientists, programmers, linguists, humanists are higher and better for achieving this goal.

Note that software considerations have been only indirectly discussed in this report. A critical review of currently available systems and an examination of their merits as well as their lacking features is further needed.

Acknowledgments

This report owes a great deal to the extensive and fruitful discussions with Alfons Geis and Conny Zuell. A lot of the misinterpretations and terminological confusions have been eliminated through their careful reading of the first drafts and their insistence on precision and

common 'terminology'. Janet Harkness, Jürgen Hoffmeyer-Zlotnik, Peter Mohler and Peter Schmidt are also acknowledged for their time and comments on the first drafts of this report. Any remaining misinterpretations and errors, however, remain my own responsibility.

References

Alexa, M. & L. Rostek (1997): Pattern concordances - TATOE calls Xgrammar. In Conference Abstracts of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing (ACH-ALLC'97), Queens University, Kingston, Canada, 1997, pp. 3-4

Barcus, F. E. (1959): Communications content: analysis of the research, 1900 - 1958. Ph.D. dissertation, University of Illinois.

Berelson, B. (1952): *Content analysis in communication research.* Free Press, Clencoe, Illinois.

Bos, W. & Chr. Tarnai (eds.) (1996*): Computerunterstützte Inhaltsanalyse in den empirischen Sozialwissenschaften.* Waxman Verlag, Münster/New York .

Carley, K. (1988): Formalizing the social expert's knowledge. *Sociological Methods and Research*, 17, pp. 165-232.

Carley, K. (1993): Coding choices for textual analysis: a comparison of content analysis and map analysis. *Sociological Methodology,* 23, pp. 75-126.

Carley, K. & M. Palmquist (1992): Extracting, representing, and analyzing mental models. *Social Forces*, 70, pp. 601-636.

Cartwright, D. P. (1953): Analysis of qualitative material. In Festinger L. & D. Katz (eds.): *Research methods in the behavioural sciences*. Holt, Rinehart, and Winston, New York, pp. 421-470.

Danowski, J. A. (1988): Organizational Infographics and automated auditing: Using computers to unobtrusively gather and analyze communication. Goldhaber, G. & G. Barnett (eds.) *Handbook of organizational communication,* Ablex, Norwood, N.J., pp. 385-433.

Deese, J. (1969): Conceptual categories in the study of content. In Gerbner *et al.* (eds.) (1969): *The analysis of communication content,* pp. 39-56.

DeWeese, L. C. (1976): Computer content analysis of printed media: a limited feasibility study. *Public Opinion Quarterly,* Vol. 40, pp. 92-100.

DeWeese, L. C., III (1977): Computer content analysis of day-old newspapers: a feasibility study. *Public Opinion Quarterly,* Vol. 41, pp. 91-94.

Dovring, K. (1954): Quantitative semantics in 18th century Sweden. *Public Opinion Quarterly*, 18, 4, pp. 389-394.

Ellis, F. S. (1968): *A lexical concordance to the poetical works of Percy Bysshe Shelley*. B. Franklin, New York.

Eltinge, E. M. (1997): Assessing the portrayal of science as a process of inquiry in high school biology textbooks: an application of linguistic content analysis. In Roberts, (ed.) (1997), pp. 159-170.

Eltinge, E. M. & C. W. Roberts (1993): Linguistic content analysis: A method to measure science as inquiry in textbooks. *Journal of Research in Science Teaching,* 30, pp. 65-83.

Ericsson, K. A. & H. A. Simon (1984): *Protocol Analysis: Verbal reports as data.* MIT Press, Cambridge.

Fischer, S.; R. Lienhart & W. Effelsberg (1995): Automatic Recognition of Film Genres. In Electronic Proceedings of ACM Multimedia 95, November 1995, San Francisco, California.

Franzosi, R. (1989): From words to numbers: A generalized linguistics-based coding procedure for collecting textual data. *Sociological Methodology*, 19, pp. 263-298.

Franzosi, R. (1990a): Strategies for the prevention, detection, and correction of measurement error in data collected from textual sources. *Sociological Methods and Research*, 18, pp. 442-472.

Franzosi, R. (1990b): Computer-assisted coding of textual data. *Sociological Methods and Research*, 19, pp. 225-257.

Franzosi, R. (1995): Computer-assisted content analysis of newspapers: Can we make an expensive tool more efficient? *Quality and Quantity*, 29, pp. 157-172.

Franzosi, R. (1997): Labor unrest in the Italian service sector: an application of semantic grammars. In Roberts (ed.), (1997), pp. 131-146.

Früh, W. (1991): *Inhaltsanalyse. Theorie und Praxis.* Ölschläger Verlag, München.

Garside, R., G. Leech & G. Sampson (1987): *The Computational Analysis of English: A corpus-based approach.* Longman, London.

Geis, A. (1992): Computerunterstützte Inhaltsanalyse - Hilfe oder Hinterhalt? In Zuell, C. & P.P. Mohler (eds.) (1992): *Textanalyse: Anwendungen der computerunterstützten Inhaltsanalyse*. Westdeutscher Verlag, Opladen, pp. 7-32.

Geis, A. & C. Zuell (1996): Strukturierung und Codierung großer Texte: Verknüpfung konventioneller und computerunterstützter Inhaltsanalyse. In Bos & Tarnai (eds.) (1996), pp. 169-191.

Gerbner, G., O. R. Holsti, K. Krippendorff, W. J. Paisley & P. J. Stone (eds.) (1969): *The analysis of communication content.* John Wiley & Sons, Inc. New York.

Goldfarb, C. F. (1990): *The SGML Handbook.* Clarendon Press, Oxford.

Goldhamer, D. H. (1969): Toward a more General Inquirer: convergence of structure. In Gerbner *et al.* (eds.) (1969): *The analysis of communication content*, pp. 343-353.

Grishman, R. & B. Sundheim (1996): Message Understanding Conference - 6: a brief history. Proceedings of COLING-96, Vol. 1, pp. 466-471, Kopenhagen, Denmark.

Grünzig, H. J. (1983): Themes of anxiety as psychotherapeutic process variables. In Minsel, W. R. & W. Herff (eds.) (1983): *Methodology in psychotherapeutic research.* Proceedings of the 1st European Conference on Psychotherapy Research, Trier, 1981. Lang, Frankfurt, pp. 135-142.

Hays, D. (1969): Linguistic foundations for a theory of content analysis. In Gerbner *et al.* (eds.) (1969): *The analysis of communication content*, pp. 57-68.

Kripendorff, K. (1969): Models of messages: three prototypes. In Gerbner *et al.* (eds.) (1969): *The analysis of communication content*, pp. 69-106.

Holsti, O. R. (1969): *Content analysis for the social sciences and the humanities*. Reading, Mass.

Hohnloser, J. H., F. Pürner & P. Kadlec (1996): Coding medical concepts: a controlled experiment with a computerized coding tool. Medical Informatics, Vol. 21, n. 3, pp. 199-206.

Iker, H. P. (1974): SELECT: A computer program to identify associationally rich words for content analysis. I. Statistical results. *Computers and the Humanities*, Vol. 8, pp. 313-319.

Iker, H. P. & R. Klein (1974): WORDS: A computer system for the analysis of content. *Behavior Research Methods and Instrumentation*, 6, pp. 430-438.

ISO (1986): *Information Processing - Text and Office Systems Standard Generalized Markup Language (SGML)*. International Organization for Standardization, ISO 8879-1986, Geneva 1986.

Kabanoff, B. (1996): Computers can read as well as count: How computer-aided text analysis can benefit organisational research. Trends in Organizational Behaviour, 3, 1996, pp. 1-21.

Kelle, U. (ed.) (1995): *Computer-aided qualitative data analysis, theory, methods and practice*. Sage, London.

Kriz, J. (1978): Methodologische Grundlagen der Inhaltsanalyse. Lisch, R. & J. Kriz (1978): *Grundlagen und Modelle der Inhaltsanalyse*, (studium)Rororo, Rowohlt, Reinbeck bei Hamburg, pp. 29-55.

Lafon, P. (1984): *Dépouillements et statistiques en lexicométrie*. Slatkine-Champion, Genève.

Lasswell, H. D. & J.Z Namenwirth (1968): *The Lasswell Value Dictionary*. New Haven.

Lasswell, H. D., D. Lerner, & I. de S. Pool (1952): *The comparative study of symbols*. Stanford University Press, Stanford.

Lazarsfeld, P. F. & A. H. Barton (1951): Qualitative measurement in the social sciences, classification, typologies, and indices. In Lerner, D. & D. Lasswell (eds*.): The policy sciences: recent developments in the scope and method.* Stanford University Press, Stanford, pp. 180-188.

Lebart, L. & A. Salem (1994): *Statistique textuelle.* Dunod, Paris.

Lienhart, R., S. Pfeiffer & W. Effelsberg (1996): The MoCA Workbench: Support for Creativity in Movie Content Analysis. Proceedings of IEEE Conference on Multimedia Computing & Systems, June 1996, Hiroshima, Japan.

Litkowski, K., C. (1997): Desiderata for Tagging with WordNet Synsets and MCCA Categories. Proceedings of the 4th SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?, April 1997, Washington, DC.

Mallery, J. (1985): Universality and Individuality: The interaction of noun phrase determiners in copular clauses. In Proceedings of 23rd Annual Meeting of the Ass. For Computational Linguistics. Chicago, 1985.

Martindale, C. (1978): The therapist-as-fixed-effect fallacy in psychotherapy research. *Journal of Consulting and Clinical Psychology*, 46, 6, pp. 1526-1530.

Mayntz, R., K. Holm & P. Hübner (1978): *Einführung in die Methoden der empirische Soziologie*. Westdeutscher Verlag (5te Auflage).

McArthur, T. (1981): *Longman Lexicon of Contemporary English.* Longman, London.

McTavish, G. D. & E. B. Pirro (1990): Contextual content analysis. *Quality and Quantity* 24, pp. 245-265.

McTavish, G. D., Litkowski, K. C. & Susan Schrader (1995): A computer content analysis approach to measuring social distance in residential organizations for older people. Paper presented in Text Analysis and Computers Conference, Mannheim, September 1995. (Paper published in *Social Science Computer Review*, 15 (2), 1997, pp. 170-180.)

Miller, G., A., Beckwith, R., Fellbaum, C., Cross, D. Miller, K. & R. Tengi (1993): Five Papers on WordNet™, CSL Report 43. Cognitive Science Laboratory, Princeton University, Princeton, NJ.

Miller, M. M. & B., P. Riechert (1994): Identifying themes via Concept Mapping: a new method of content analysis. Presented to the Theory and Methodology Division, Association for Education in Journalism and Mass Communication Annual Meeting, August 1994. Electronic version: http://excellent.com.utk.edu/~mmmiller/pestmaps.txt

Morris, R. (1994): Computerized content analysis in management research: A demonstration of advantages and limitations. *Journal of Management,* 20, pp. 903-931.

MUC-3*: Proceedings of the Third Message Understanding Conference (MUC-3)*, August 1993. Morgan Kaufmann Publishers, San Diego, CA, USA.

MUC-5*: Proceedings of the Fifth Message Understanding Conference (MUC-5)*, August 1993. Morgan Kaufmann Publishers, San Fransisco, CA, USA.

Nazarenko, A., B. Habert & C. Reynaud (1995): "Open response" surveys: from tagging to syntactic and semantic analysis. In Proceedings of JADT (3rd International Conference on Statistical Analysis of Textual Data), Vol. II, pp. 29-36, Rome, Italy, 1995.

Polanyi, L. (1985): *Telling the American story: A structural and cultural analysis of conversational storytelling.* Ablex, Norwood N.J.

Pool, I. de Sola (ed.) (1959): *Trends in Content Analysis*. University of Illinois Press, Urbana, Illinois.

Roberts, C. W. (1989): Other than counting words: a linguistic approach to content analysis. *Social Forces*, 68, pp. 147-177.

Roberts, C. W. (1997a): Semantic text analysis: On the structure of linguistic ambuguity in ordinary discourse. In Roberts (ed.) (1997b), pp. 55-78.

Roberts, C. (ed.) (1997b): *Text analysis for the social sciences: methods for drawing statistical inferences from texts and transcripts*, Lawrence Erlbaum Assoc. Publishers, Mahwah, N.J.

Roberts, C. W. & R. Popping (1996): Themes, syntax and other necessary steps in the network analysis of texts: a research paper. *Social Science Information* 35(4), pp. 657-665.

Sacks, H. (1972): On the analysability of stories by children. In Gumprez, J. & D. Hymes (eds.): *Directions in Socio-linguistics*. Holt, Rinehart & Winston, New York, pp. 329-345.

Sperberg-McQueen, C.M. & L. Bournard (eds.) (1994):*Guidelines for the Encoding and Interchange of Machine-Readable Texts (TEI P3).* Chicago & Oxford, April 1994.

Stone, P. J. (1997): Thematic text analysis: new agendas for analyzing text content. In Roberts, C. (ed.) (1997), pp. 35-54.

Stone, P. J. (1969): Improved quality of content analysis categories: computerized-disambiguation rules for high-frequency English words. In Gerbner *et al.* (eds.) (1969), pp. 199-221.

Stone, J. P., Dunphy, D. C., Smith, M. S. & D. M. Ogilvie (1966): *The* General Inquirer: *a computer approach to content analysis.* MIT Press, Cambridge, MA.

Tesch, R. (1990): *Qualitative research: Analysis types and software tools.* Falmer, New York.

van Dijk, T. (1972): *Some aspects of semantic grammars.* Mouton, Paris.

van Dijk, T. (1988): *News as discourse*. Lawrence Erlbaum Assoc. Publishers, Hillsdale, N.J.

van Lehn, K. & S. Garlick (1987): Cirrus: An automated protocol analysis tool. In Proceedings of the 4th International Workshop on Machine Learning, edited by P. Langley. Morgan-Kaufman, Los Altos, California, pp. 205-217.

Weber, R. (1990): *Basic content analysis*. Sage Publications, 2nd edition, Newsberry Park, California.

Weitzman, E. B. & M. B.Miles (1995): *Computer programs for qualitative data analysis*. Sage Publications, Thousand Oaks, CA.

Wilson, A. & P. Rayson (1993): The automatic content analysis of spoken discourse: a report on work in progress. Electronic reference:
http://www.comp.lancs.ac.uk/computing/research/ucrel/papers/war93.txt.

Zuell, C., Mohler, P. P. & A. Gleis (1991): *Computerunterstützte Inhaltsanalyse mit TEXTPACK PC*. Gustav Fischer Verlag, Stuttgart.

Zuell, C., P. W. Weber & P. P. Mohler (1989): Computer-aided text classification for the social sciences: The General Inquirer III. Mannheim, Germany: ZUMA, Center for Surveys, Research and Methodology.